Stony Brook University

Quantitative Approaches for Deconvolving the Multiple Contributions of Primary Structure to Protein Fitness

A Dissertation Presented

by

Loretta Au

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2014

Stony Brook University

The Graduate School

Loretta Au

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

David F. Green – Dissertation Advisor
Associate Professor & Graduate Program Director, Department of Applied Mathematics and Statistics

Robert C. Rizzo – Chairperson of Defense
Associate Professor, Department of Applied Mathematics and Statistics

Thomas MacCarthy – Third Member
Assistant Professor, Department of Applied Mathematics and Statistics

Steven Glynn – Outside Member
Assistant Professor, Department of Biochemistry and Cell Biology

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Quantitative Approaches for Deconvolving the Multiple Contributions of Primary Structure to Protein Fitness

by

Loretta Au

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2014

Proteins from the same family often share many structural and functional motifs. Variations in primary structure, however, allow a single protein family to modulate a broad range of biological processes. Several methods can be used to identify sequence conservation, but complementary mutagenesis experiments are often needed to understand the multiple roles that a given amino acid may have in maintaining overall fitness. Limitations related to financial and temporal costs generally constrain these experiments to smaller proteins or peptides, or to only partial sampling of sequence spaces, and we have devised a computational protocol for large-scale mutagenesis to circumvent these obstacles using a G-protein heterotrimer as a model system.

The dead-end elimination and A* search (DEE/A*) algorithms are typically used to find a small number of sequences that may enhance the current level of fitness or introduce novel functions into a protein. These algorithms were adapted to find all low-energy sequences and their corresponding structures, allowing us to disentangle protein fitness, defined here as a combination of structural stability and binding interactions. We demonstrate the effectiveness of DEE/A* in capturing the biophysical features of amino-acid substitutions, and in quantifying the extent that individual positions are affected by mutagenesis, based on all low-energy single mutants. A modified version of this protocol was also used

to explore double and triple mutants in the context of the β subunit in the G-protein heterotrimer, a representative repeat protein from the β-propeller family. Repeat proteins are subject to the same experimental challenges as non-repeating ones, but with the addition of having to properly define corresponding repeats in primary structure; DEE/A* is leveraged here to identify patterns of important interaction using a structure-based approach, and reveal a deeply connected interaction motif.

To my parents, for their unconditional love and support.

Contents

List of Figures

List of Tables

Acknowledgments

There are many people in my life for whom I have been and remain grateful to have encountered. Without their help, many of my achievements and successes thus far, including the completion of this thesis, would not have been possible.

First and foremost, my advisor, Dr. David Green, has provided me with immeasurable support and guidance over the last several years. His mentorship, both in scientific research and academic teaching, will be invaluable to my career, and I remain thankful for his patience, compassion and friendship throughout my time in his research group. Not only did he help me develop my sense of independence and confidence as a scientist, but David has been supportive of my other interests as well. This includes teaching and mentoring younger students in STEM, studying abroad in Singapore, and also my many attempts to moonlight as a photographer during lab events or even at major conferences. On top of all this, I am grateful that David is a very understanding and diplomatic person, who has a good sense of humor, because this is probably how I got away with occasional shenanigans in the lab.

I would also like to thank my committee chair, Dr. Robert Rizzo, for providing me with guidance, helpful insight and advice on how to strengthen my research, and become a better scientific writer. I would also like to thank my previous and current thesis committee members, Dr. Joshua Rest, Dr. Hongshik Ahn, Dr. Thomas MacCarthy and Dr. Steven Glynn, for their useful advice and their help in polishing some of the details of this research.

I would like to thank my colleagues in the Green and Rizzo research groups, for their support and friendship over the years. In particular, Dr. Vadim Patsalo and Dr. Yukiji Fujimoto have been exceptionally helpful during our time together in Green lab. I knew nothing when I first started, but their encouragement and assistance, fortunately, helped change that.

Through East Asia and Pacific Summer Institutes (EAPSI) fellowship program, sponsored by the National Science Foundation (OISE-1209984), I was able to study and conduct research in Singapore in 2012. It was an enriching and productive experience, and I thank

Dr. Christopher Hogue and Dr. Lisa Tucker-Kellogg for their time and guidance during my stay at the National University of Singapore. I would also like to thank my fellow EAPSI cohort members and summer labmates for the countless adventures, memories and experiences that we all shared together.

Saving the best for last, I would like to thanks my family, especially my parents, and my friends. Mom and Dad both sacrificed a lot throughout my childhood, so that my brother, Randy, and I could have the opportunities that we did and continue to presently have. Their dedication as parents instilled a good work ethic and many good values in me, and have helped strengthen my sense and understanding of what is truly important in life. I would also like to thank Randy for being my first childhood friend; growing up with an older brother like him had a very postive impact on me, and sparked my curiosity about science and mathematics rather early, which has inevitably steered me towards the career and pursuits that I have today. Finally, I would also like to thank all of my friends, both whom I met during my time at Stony Brook and back home in New York City, for being the supportive network of people that I could always turn to and count on for anything.

Chapter 1

General Introduction

Protein primary structure largely dictates how a protein can stably fold into a three-dimensional conformation. Consequently, the resulting tertiary structure directly affects protein interaction specificity with different ligands, ability to bind and associate with other proteins, allosteric mechanisms involved in structural changes, and other roles essential for modulating biological processes. Sequence–function relationships are usually complex for this reason: all amino acids must maintain an appropriate level of molecular interaction to collectively satisfy these requirements, so multiple roles are often adopted by every amino acid in the protein to maintain overall fitness. Disentangling these roles can directly provide insight on how sequence variability yields functional diversity in proteins, a fundamental step towards understanding how proteins from the same family are able to modulate a broad range of biological pathways, despite sharing several structural and functional motifs. Although the relationship between primary structure and function is well-established, it is a complex one, and how the multiple aspects of fitness can be deconvolved remains an open question in biology. In this dissertation, a combination of computational and quantitative methods will be used to address this problem, to compute the fitness landscape of a G-protein heterotrimer and further investigate features of structural stability that enable β-propeller stability. The methodology and analytical techniques used can be applied to any protein

system, as long as sufficient structural information is available for it.

1.1 Heterotrimeric G-proteins

Many cellular pathways in eukaryotic organisms are modulated by heterotrimeric G-proteins, a family of proteins belonging to the GTPase superfamily that are named after their ability to bind guanine nucleotides.[1, 2] G protein heterotrimers are involved in countless signalling pathways: they have an important role in metabolic processes, managing transcription factors, mechanisms for cell motility, arrangement of the cytoskeleton, cell-cycle regulation, and even in modulating eletrostatic potential, as in ion channels or transporters.[3, 4] Due to their high level of involvement in regulating biological processes, interruptions in the signalling pathway can cause the onset of diseases such as cancer, heart disease, blindness or hormonal disorders.[5, 6, 7, 8] Mutations in the heterotrimer can block receptor signals or enhance them, and these imbalances will disrupt proper cellular function.

Four main families have been identified in G protein heterotrimers, based on sequence similarity of the α subunit: G_s (stimulatory), $G_{i/o}$ (inhibitory), G_q/11 and G12/13).[9, 3] The human genome is known to express at least 23 α, six β and 11 γ subunits,[9] and a fully formed protein complex requires one of each. Signal transduction begins when a ligand—a neurotransmitter or hormone, for example—is bound to a heptahelical G-protein coupled receptor (GPCR) at the plasma membrane (**Fig. 1.1**): a conformational change is induced in the GPCR that, in turn, causes a conformation change in the α subunit, favoring the binding of GTP and a magnesium ion (Mg^{2+}) over GDP.[10, 11] In order to properly interact with the G-protein coupled receptor, a number of post-translational modifications are required, namely palmitoylation and myristoylation on the N-terminus of the α subunit and C-terminus of the γ subunit, so that attachment to the plasma membrane remains possible.[12, 8] Structural changes in the GTP-bound α subunit favor dissociation from the $\beta\gamma$-heterodimer, allowing each component (the active α subunit and heterodimer)

to modulate different downstream effectors.[13, 7] Gα, being closely related to the Ras protein family, can dephosphorylate GTP in the ligand-binding site; by cleaving the terminal phosphate group of GTP, the α subunit returns to its original, inactive GDP-bound state, and can act as a negative regulator of heterodimer activity by re-establishing itself with G$\beta\gamma$, and repeating the cycle when another ligand is bound.[14] The interaction in the coiled-coil region of the $\beta\gamma$-heterodimer is highly stabilizing, and the two subunits are believed to remain associated together throughout the pathway.[15] This classical activation pathway is widely accepted, but some experiments have also suggested that subunit rearrangement (rather than complete dissociation) may occur instead for some heterotrimers.[16, 17, 18]

1.1.1 Structural features of Gα and $\beta\gamma$-heterodimer

Conservation in primary structure can vary from one subunit to the next, with the highest in Gβ: approximately 80% sequence homology is found across all Gβ subunits, except Gβ_5, which shares about 50% identity with all the others. Next, Gγ subunits have at least 50% identity between all members, except Gγ_1.[19, 13] In Gα, there is at least 40% sequence identity between the four major classes of α subunits; within any of these classes, however, as much as 80% of the sequences are similar, as seen in Gα_q.[20]

A number of structural motifs are shared by proteins within each subunit class despite variation in primary structure (**Fig. 1.2**). Gα is closely related to the Ras family of small, soluble G-proteins, and is primarily responsible for catalytic activity in the α-subunit; a helical domain, comprised of six α-helical bundles, and the GTPase forms a binding cleft between them for the phosphorylated nucleotide. The N-terminus extends away from the α-subunit and can interact with a portion of the Gβ-binding interface.[21]

The $\beta\gamma$-heterodimer is thought to remain associated, unless it is exposed to denaturing conditions, since many mutually stabilizing interactions exist between them, including a coiled-coiled motif between their N-termini. Gβ is a member of β-propeller proteins, a repeat protein family: repeating regions of this protein family are comprised of approximately 40–60

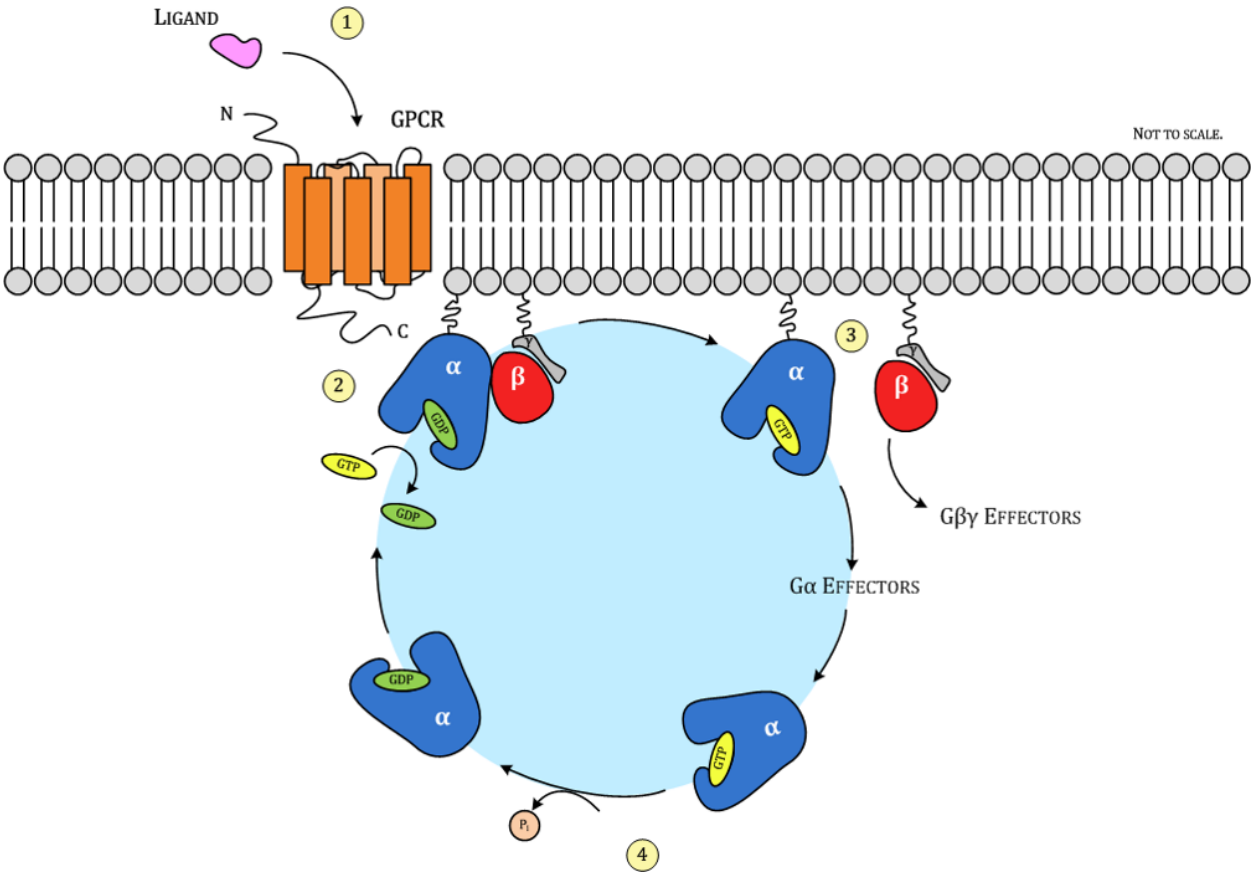


Figure 1.1: In the inactive state, heterotrimeric G-proteins are bound to the plasma membrane, near the C-terminus of a G-protein coupled receptor (GPCR, orange). (1) A ligand binding to the GPCR will induce a conformational change in Gα; (2) binding GTP becomes more favorable, so GDP is released. In the classical activation pathway, (3) GTP-bound Gα will dissociate from the $\beta\gamma$-heterodimer, and each can modulate different biological pathways. (4) Dephosphorylation of GTP will return Gα to its inactive state, and the cycle can repeat.

amino acids that form anti-parallel β-sheets, including a signature tryptophan and aspartate. Four to eight propellers are formed and arranged in a toroidal fashion, allowing the N- and C-terminal repeats to interact. Nearly 1% of all proteins are β-propeller proteins, often

involved in binding interactions.[22]

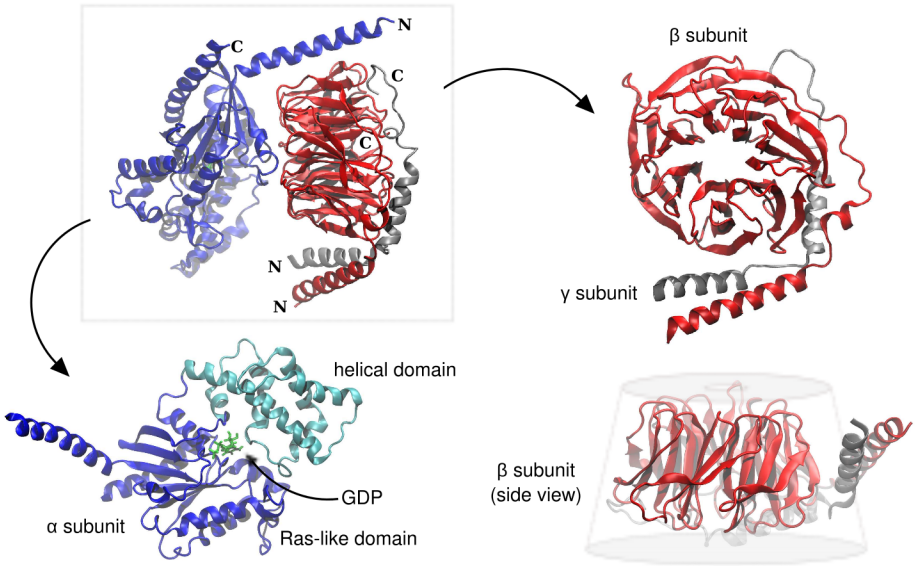


Figure 1.2: A complete heterotrimer is comprised of an α (blue), β (red) and γ subunit (gray). In the inactive state, GDP is bound between the Ras-like and helical domains on Gα. The γ subunit is stabilized by the coiled-coiled motif at the N-terminus of the heterodimer. The β subunit has seven repeating regions, and is geometrically disc-like in shape.

Unlike its partner, Gγ is significantly smaller than either of the other two subunits: it is a sequence of (approximately) 60 residues, forming two sequential α-helices and a prominent C-terminal domain with a CAAX motif, where isoprenylation at cysteine may be possible. Post-translational modification with this hydrophobic molecule is thought to facilitate heterodimer-binding to the plasma membrane.

1.1.2 Associated conditions and disease

Over a thousand GPCRs can be expressed in the human genome, and at least 200 ligands, such as peptides, ions, lipids, odorants and more, are known for them.[23] How G-protein specificity is achieved for these receptors, however, remains unclear,[23] and subunit specificity in forming a biologically functional heterotrimer is also an open problem. Furthermore, there are exceptions to classical activation pathway: some G-protein heterotrimers can be biologically active without any interaction with a G-protein coupled receptor, and this be-

havior has been found in the G_i family.[24, 25] Regardless of the origins of signaling, given the high prevalence of G-proteins in modulating biological pathways, many aspects of human physiology can become affected when problems in signaling arise. Cardiovascular diseases related to the autonomic control of cardiac muscles can lead to heart failure, for instance, and changes in cardiac tissue can cause complications due to smooth muscle tone or myocardial hypertrophy.[26, 27] Several known ligands for GPCR are hormones, and thus a number of endocrine disorders are related to heterotrimeric G-proteins, such as carbohydrate and lipid metabolism.

There are three common mechanisms that can affect the α subunit and cause disease: (1) GTPase activity may be disrupted, consequently prolonging the active state of Gα; (2) G-protein subunits may be genetically inactivated, so the protein is not expressed; and (3) abnormal signal initation may occur.[28] Several lifestyle factors are associated with many of the known diseases for the G-protein heterotrimers mentioned earlier, but two are directly linked to mutations in Gα alone, due to inactivating $G_s\alpha$ mutations.[29] Albrights' hereditary osteodystrophy (AHO) was the first Gα disease to be characterized.[30] Patients with AHO have varying symptoms, depending on whether the maternal or paternal allele was affected. Resistance to parathyroid hormone, a regulator of serum calcium and phosphate, is a major symptom, and this is directly related to various types of pseudohypoparathyroidism[5, 31, 29] McCune–Albright syndrome is the other condition that is a consequence of mutations in Gα during the early stages of development (instead of genetic inheritance), in which signaling is prolonged due to disruption in GTPase activity in the α subunit.[31] Classical symptoms for this disease include hyperpigmentation, if melanocytes are affected, in which patients can have patches of darker skin tone and tumors or fibrous dysplasia, a condition in which skeletal structure is not properly formed, replaced by tissue instead, if the bone marrow is affected.

1.2 Exploring protein sequence–function relationships

A number of approaches for measuring the relatedness of different proteins exist, but complementary experiments are often required to understand the connection between primary structure, structural stability and protein function. Several algorithms, for example, can be used to measure sequence conservation, but further investigation is needed to interpret reasons for conservation at different positions in primary structure. The most direct approach for deconvolving the underlying roles of primary structure is to replace wild-type amino acids, then measure changes in structural stability, binding or any other aspect of protein fitness; many techniques like alanine scanning, guided mutagenesis studies and directed evolution experiments have focused on identifying important amino-acid substitutions that can alter protein fitness. However, with these methods, the temporal and financial costs associated with experimentation can become too great, or the protein system of interest is too large, so only a limited region of protein sequence space can be explored as a compromise: an average-sized protein may have around 350 amino acids, for example, and almost 7000 single-mutant sequences are possible. Without fully exploring all of these mutations, there can be constraints or unavoidable biases in interpreting the mutagenesis results. As a mechanism to overcome these challenges, a feasible alternative was found in adapting well-established computational protein design algorithms to efficiently search through protein sequence and structural spaces simultaneously for large-scale *in silico* mutagenesis. A brief survey and comparison of existing methods are presented here to clarify how the algorithms were chosen for computational mutagenesis.

1.2.1 Computational sequence-based methods

Several techniques in comparative sequence analysis exist, and are helpful in addressing how highly or poorly conserved a residue may be within a given protein family. Conservation scores may be computed for any given set of protein sequences, and structural information

is not needed for them, providing major advantages to using these techniques, but many challenges can arise in properly interpreting conservation scores within the context of biology. Typically, conservation is measured based on how often a specific amino acid is found at an aligned position, and if amino-acid variation exists for the position, well-established substitution frequencies of one amino acid for another are used to address this disparity. Similarity matrices are constructed by measuring the frequency of substitution for known protein sequences either for a specific region (as in BLOSUM, *BLO*cks *SU*bstitution *M*atrix) or without any restriction after alignment (as in PAM, Point Accepted Mutation, and Gonnet matrices.)[32, 33, 34] A number of algorithms can measure the relationship between sequences in terms of conservation scores,[35, 36, 37, 38] which generally fall on the range of 0–100%. Having a score near zero indicates that amino acids are largely disparate—side chains have nothing in common biochemically at that position and the position may be irrelevant for biological function. On the other hand very high, near-perfect scores are assigned to positions that may have an essential role in fitness, and identical or functionally similar amino acids are found at that position. The definitions at either end of conservation are easy to interpret, but the scores in between are less obvious to understand within a meaningful biological context. In all cases, though, additional analysis is required to fully understand how a specific side chain contributes to protein fitness.

There are additional limitations in using conservation scores to directly elucidate sequence–function relationships, and this lies in how the protein sequences are chosen for an alignment. Most obviously, the number of sequences used may be too few, and thus meaningful conclusions might not be possible for statistical reasons. Secondly, if the sequences are too different in sequence length, gap insertions for different regions of the protein will be necessary, but how these place-holders should be incorporated may be unclear: their placement can vary from one set of proteins to another, and having gap insertions can alter the distribution of alignment scores when its positioning is changed. Phylogenetic relationships between sequences will also influence interpretation:[39] high amino-acid similarity at several

positions between two completely unrelated proteins may suggest some common mechanism in achieving protein stability or function, since low sequence identity is expected between an evolutionarily distant pair. On the other hand, finding a few poorly conserved positions between two, more closely related proteins (expected to have high conservation throughout most of the sequence) may be a possible indication that the positions have a role in interaction specificity. The appropriate interpretation would also depend on how much further a a given protein sequence has already evolved or may potentially continue to do so.

Addressing amino-acid co-dependencies. Although sequence alignment algorithms can measure how consistently the same amino-acid or biochemically similar ones are found at a given position, the calculations are made without consideration of coupled effects between pairs of amino acids. This shortcoming has been recognized, and alternatives using statistical analysis to identify side-chain co-variation and dependencies have been developed.[39] Statistical coupling analysis (SCA) has demonstrated that allosteric pathways can be derived from comparative sequence analysis of protein domains.[40, 41, 42, 43] Amino-acid sequences from the same protein family are first aligned, then subsets of these sequences are chosen based on features that are shared between them (for instance, a highly conserved position that is found within the subset, but not the entire family.) A second alignment is performed on this subset, and then the amino-acid conservation at all other positions (not the one used as a selection criteria) is compared with the distribution found in the original alignment. Any significant changes between the two sets of aligned sequences are an indication of co-dependency.

Alternatively, co-dependency can also be determined using quantitative approaches based on mutual information measurements. Mutual information is an entropy-based metric between two or more variables, and it describes their mutual dependence on each other. By itself, mutual information can be computed over all pairs of amino acids in a sequence alignment to identify co-dependent pairs. Mutual information is also included in direct-coupling analysis (DCA), which focuses on separating pairs that have evolutionary interdependence

from indirectly coupled ones, such as spatially constrained protein contacts.[44, 45] Monte Carlo or heuristic algorithms can be used to evaluate directed coupling for a given sequence alignment; the approximations made by the latter can be transformed into mutual information scores, and comparisons between each pair in the multiple sequence alignment are made to distinguish one type of coupling from the other. Although DCA has been used to identify and even recapitulate important contacts in protein tertiary structure,[46] significance thresholds for mutual information can be difficult to define. Similar to conservation scores in a sequence alignment, identifying important computed values is often statistically based, so the connection between these numbers and biological context may be weakened as a result: biology is rarely a function of numerical cutoffs, and how statistical significance is defined directly affects the interpretation of DCA. Still, the application of DCA in identifying different types of coupled contact is important,[47] but it is also possible that performance may vary for different protein families, if a sufficiently large multiple sequence alignment cannot be established.

Evolution-based approaches have tried to reconstruct the genetic ancestral history of a given protein to gain insight on how mutations can have beneficial or adverse effects on overall fitness; this can offer an important perspective on how amino-acid variation can withstand various selection pressures, in addition to deriving predictions on how existing genes may continue to change.[48, 49, 50] This approach has heavily relied on statistical analysis and computational simulations of sequence variation in the past, but more recent developments in protein expression techniques have made experimental validation feasible. Potentially important protein sequences can be derived by following an evolutionary pathway, which may yield specific sequences, or a limited set of them, for experimentation. In considering the required adaptations that are made at the genetic level, ancestral gene reconstruction will inevitably have sampling biases: protein sequences can only be evaluated if they have overcome various evolutionary pressures, and thus it remains unclear whether a protein has failed to adopt alternative mechanisms for thriving in an environment (an adaptive solution

is never possible for the protein), or if it has not evolved long enough to find them (an adaptive solution may be possible, but not yet found.) Genetic relationships that enable survivorship, on the other hand, is already rather complex, but despite this, information derived from ancestral gene reconstruction can still give insight on how fitness requirements may be acquired.

1.2.2 Generating protein sequence libraries

Large protein sequence libraries can be generated by substituting wild-type amino acids, and observing changes in phenotype to determine whether native interactions were essential to a specific aspect of fitness. In high-throughput phage display, protein sequences are expressed onto phage, eluted according to binding properties or other pre-determined features of protein fitness, and the remaining sequences are then amplified.[51, 52, 53] Applying directed evolution experiments to the initial generation of protein sequences help define the functional role of specific amino-acid changes, and these allowed variations are based on the chosen fitness criteria. In this way, very diverse amino-acid sequences can be generated and expressed as proteins, but only a limited number of fitness aspects can be explored. By having selection pressures filter out sequences between successive steps, the final set of surviving sequences are expected to have biases towards the pre-selected aspects of fitness, and it becomes less clear how other attributes may have been affected; improvements for unbiased sampling could be made if mutagenesis were performed systematically to all wild-type amino acids to all other possible substitutions. Recognizing the importance of this, efforts in systematic mutagenesis have been made in a modestly sized protein: single mutants have been constructed in barnase, a 109-position ribonuclease, to identify positions that can tolerate side-chain substitutions and still retain catalytic function.[54] The feasibility of doing the same in larger protein systems, however, remains unclear.

The intentions of alanine scanning are similar to full-scale mutagenesis experiments in exploring an unbiased protein sequence library, but the number of protein sequences is

substantially more manageable: alanine-scanning techniques measure changes in fitness after wild-type side chains are replaced by alanine.[55, 56] Due to its size and non-polar properties, alanine is an excellent candidate for determining the role and significance of some native interactions; mutations to alanine should result in some loss of stability and probably in function if a key interaction is lost, for example. While this type of substitution is relatively straight-forward, interpretations may be less obvious in densely packed, hydrophobic regions, where replacement to a smaller, aliphatic residue may be desirable. Still, alanine scanning greatly reduces the size of protein sequence space required for meaningful analysis, and computational methods have also been developed to imitate this procedure with greater efficiency.[57, 58]

1.2.3 Computational structure-based methods

Computational protein design. Approaches in computational protein design (CPD) are structure-based methods that continue to make progress in answering how amino acid sequences are able to influence protein structure and function.[59, 60, 61] In understanding how amino-acid sequences are related to protein tertiary structure, it is possible to think about CPD goals in two ways: (1) focus on modifying an existing sequence to alter some aspect of its current level of fitness or (2) design a *de novo* sequence that can adopt a target fold or function. A number of successes in CPD have been in redesigning existing proteins to enhance features such as stability or specificity, often by finding novel sequences that can adopt a target fold or function that would have been difficult to achieve using sequence-based methods.[62, 63, 64]

Different types of algorithms can be used to address the sequence–function relationship, and fall into two categories: stochastic methods and deterministic ones.[65] Stochastic algorithms rely on decision making that is rarely identical from one simulation to the next. Genetic algorithms, for instance, attempt to mimick evolutionary events: mutations may happen to the sequence, one amino acid at a time, or a cross-over event between sequences

may occur.[66] Alterations are usually made according to a uniform distribution, and sequences are evaluated according to a set of selection criteria that defines fitness. Monte-Carlo based methods, on the other hand, typically accept or reject a sequence based on a Boltzmann distribution (as seen in the Metropolis criterion for searching state space [67]); this probabilistic approach biases sampling towards changes that are progressive—if the previous sequence is more favorable than the one currently found by the algorithm, the current sequence will be discarded from the solution space. Even though Monte-Carlo methods have preference for energetically favorable sequences, only a limited portion of low-energy sequences are typically explored, and thus some disparity is expected between the found sequences and the complete set of all possible low-energy protein sequences. This strategy for global optimization is implemented in Rosetta,[68, 69] in which protein conformations are assembled using fragments of known proteins according to sequence information that is provided. Although finding the same results repetitiously might not be possible using stochastic models, the operations used towards converging to a solution can often help to avoid becoming trapped in local energy minima and to avoid sampling all possible combinations in a given conformational space. In fact, the popularity of Monte Carlo sampling originated in its effectiveness at finding energy minima, when multiple solutions may be possible, [70] and so a variety of Monte-Carlo based approaches are available, often to improve sampling resolution or the convergence time.[71, 72]

In contrast, deterministic models may converge less quickly, but offer reliability in reaching the same solution every time that the same simulation is run.[59] Self-consistent mean-field approaches formulate the sequence–structure problem using mean-field theory: the allowable choices on a protein backbone are organized into a matrix, and the potential energies from interaction in the native conformation is computed, so that favorable states can be distinguished from less favorable ones.[73, 74] All possible conformations have a probability of occurrence based on this potential energy, and higher probabilities are assigned to the states that are more favorable. Potential energy is then recomputed with this new prob-

ability distribution for the same protein conformation; updated probabilities are assigned again, and this iterative process continues until a converged solution or set of solutions has been reached. Dead-end elimination requires a similar setup in which rotamer states need to be defined for a given protein conformation. Similar to self-consistent mean field, a mechanism is required to effectively consider all possible choices, but dead-end elimination will reach convergence by discarding improbable solutions along the way. Typically, a dead-end elimination algorithm will make energetic comparisons between unique pairs of possible rotamers; depending on how each rotamer in the pair interacts with the protein backbone and all neighboring amino acids, the more favorable option between the two is kept, and this procedure continues until all states that are incompatible with a low-energy structure have been pruned from the initial set of amino acid and rotamer choices.[75, 76, 77, 78, 79] Different combinations of the remaining rotamers are searched with a heuristic function to identify low-energy sequences and their corresponding structures simultaneously. Additional details of implementation and variation are discussed in **Chapter 2**.

1.3 Specific aims of research

To disentangle and quantify the multiple aspects of fitness that amino acids may have, a computational protocol was developed here to efficiently find all low-energy sequences for a G-protein heterotrimer, $G_i\alpha_1\beta_1\gamma_2$, and evaluate the extent that every amino acid is involved in protein fitness, defined here as a combination of structural stability and binding interactions. Existing methods for understanding protein sequence conservation and its relationships with structural stability and binding interactions were first surveyed (**Chapter 1**). This comparison includes methods that require biological experiments, and also relevant areas of computational biology that have been used to address how sequence and function are related; the advantages and shortcomings of each approach are discussed. After assessing the available options, the dead-end elimination and A* search (DEE/A*) algorithms

were chosen to systematically mutate the heterotrimeric G-protein (**Chapter 2**); the search for all single mutants was limited to only low-energy sequences, with the assumption that high-energy ones are unlikely to be found in biology. For every mutant, a corresponding protein conformation was also evaluated, and the differences between all mutations relative to the wild type were compared. The effect of each mutation at every position was mapped out, so that regions of mutational sensitivity and robustness could be found; these results were compared against alternative computational methods that evaluated side-chain substitutions, and the frequency of amino-acid substitutions found in DEE/A* was compared with expectations derived from amino-acid similarity matrices. A strong correlation was found between the biophysical interactions captured by DEE/A* and these similarity matrices, and the resolution of mutational effects in understanding structural context is an improvement over other existing computational methods. This was an indication that DEE/A* was an appropriate choice for simulating and evaluating mutations to wild type for understanding how substitutions affected specific aspects of protein fitness.

In addition to this, the primary advantage of using DEE/A* is in avoiding many obstacles that techniques in comparative sequence analysis have, which is commonly used to understand conservation. In particular, the correct approach for aligning sequences can be ambiguous in some protein families, as demonstrated by proteins with tandem repeats: the repetitive nature of repeat proteins can complicate the interpretation of multiple sequence alignments. Primary structure can be aligned according to the entire sequence for the repeat protein, or the sequence may be truncated so that each corresponding repeat is aligned with each other; the latter provides greater multiplicity of conservation data than the former. In both, the signature residues will maintain a strong signal in the alignment for being highly conserved, and sequence variation at other positions will be harder to interpret, a problem that is not limited to repeating proteins. However, when the number of repeats are not the same, how repeats should be aligned against each other becomes unclear. There is also the possibility of evaluating repeats by defining where the boundary of each repetitive region

might be, but the size of each repeating region in primary structure can be highly variable or it may not be directly compatible with what is seen structurally. In **Chapter 3**, DEE/A*-based protocol is extended to address how patterns of interaction can be determined in a β-propeller repeat protein, the β subunit of the G-protein heterotrimer, and how mutational robustness is achieved; the symmetric nature of this protein also motivated questions in defining coupled interactions for a given protein system.

Related to the coupling of protein interactions is the concept of epistasis, how a mutant sequence can only be found if specific background mutations are present. The extent that DEE/A* is able to capture some of these effects, and how they are relevant in the context of the β subunit was of interest (**Chapter 4**). The efficiency of DEE/A* for mutagenesis was pushed further, so that the analysis of single mutants in the context of double and triple mutants became possible. In the hypothetical 350 amino-acid protein, though, there are over 61,000 and 7 $\times 10^6$ unique pairs and triplets, respectively; this yields about 2.4×10^7 and 5.6×10^{10} double and triplet mutants, respectively, and structural spaae will also grow. Even with the devised protocol with DEE/A*, it was not possible to explore all low-energy double and triplet mutants within a reasonable amount of time, with the available computational resources. Also, most positions in these pairs and triplets would be very far from each other, and were expected to behave as single mutants. As a compromise, pairs and triplets were chosen based on distance cutoffs, and probabilistic approaches were derived to determine whether amino-acid co-dependencies could be found in the context of structural stability.

Large-scale mutagenesis using DEE/A* was demonstrated in this dissertation to be a useful tool for disentangling the behavior and interactions of individual amino acids as a part of the entire protein. With this computational approach, intuition and perspective for mutational robustness in a given protein system could be established: changes to wild type may have minimal or dramatic effects, and these can be separately distinguished for structural stability and for binding interactions. Some adaptations to the computational protocol were helpful, when multiple mutations were allowed in a sequence at a given time. Still,

the protocol can be further devised so that protein sequence space could be explored more fully. This method can potentially assist other protein scientists in making decisions about appropriate protein systems to use in experimentation, and also the types of mutagenesis studies that can be carried out for it. Although DEE/A* has its strengths and utility, it is also understood that the computational protocol cannot perfectly capture all biophysical relationships found in proteins. Conclusions that can be drawn from this data, its implications to protein evolutionary pathways, and areas for continued improvement are discussed in **Chapter 5**.

Chapter 2

Efficient large-scale computational mutagenesis

2.1 Introduction

Sequence–function relationships are difficult to understand in an unbiased manner, due to constraints in many current methods which often rely on analyzing amino-acid sequences that have already been filtered by various selection pressures: either through natural or directed evolution, successive generations of protein sequences must satisfy a pre-defined definition of fitness.[80, 81] Computational protein design algorithms provide accessibility to different areas of protein sequence space, including those that have not circumvented evolutionary pressures or with mutations that have not become fixed into protein evolutionary history. These algorithms have been developed over the last several decades to address how primary structure can be strategically optimized, often as *de novo* sequences, but still remain stable and functional.[62, 82, 83, 84, 85, 86, 87] Wild-type sequences can be designed to improve affinity towards a binding target, for example, or the current level of specificity with this and other binding partners can be modified; similarly, enzymatic functions may be re-designed so that catalysis is enabled under a broader or more narrow set of environmental conditions.

In methods like homology modeling or protein sequence threading, the relationship between primary structure and a possibly related target protein fold can be addressed too, but the predictive ability can be fickle at times: protein folds are generally well-conserved, even when sequence identity is relatively low,[88, 89] but it is also true that a number of novel folds may be possible, even when only a few amino acids are varied within the sequence.[90, 91, 92] By utilizing methods in computational protein design, very systematic approaches can be developed to evaluate alterations in tertiary structure, using known protein backbone conformations as a starting point.

The dead-end elimination and the A* search (DEE/A*) are well-established algorithms in computational protein design that have been successful in designing proteins for enhancing existing functions or creating novel ones. [62, 83, 84, 85, 86, 87] DEE is a deterministic approach that compares different amino acids, or their rotational isomers, by assessing how well they fit into a given structural context, through energetic evaluation.[75, 76, 77] A starting protein conformation is required in addition to a rotamer library, which will list possible amino acids and rotational isomers that can be used to replace side chains on the initial backbone. This yields a well-defined conformational space for the algorithm to perform an energetic assessment of how substitutions will alter the potential energy of the molecule.

The potential energy of a given protein sequence is computed as the sum of the backbone energy, the interactions of position i with the protein backbone, and also how i interacts with neighboring side chains at position j (**Eq. 2.1**):

$$E = E_{backbone} + \Sigma_i E(i) + \Sigma_i \Sigma_{i<j} E(i,j), i<j \tag{2.1}$$

By construction, it should be noted that pairwise decomposition is necessary in the energy calculations, or else the algorithm will fail. The self- and pairwise interaction terms are used to evaluate and eliminate the worser performing orientations: if rotamers r and t are both possible options at position i, given that rotamer s is inserted at position j, the energy difference between r and t can be compared against a target energetic cutoff, ε_{cut}. The

performance of r and t is evaluate for all rotamers that are found at neighboring position$j$, before deciding whether r or t is more favorable (**Eq. 2.2**):

$$[E(i_r) - E(i_t)] + \Sigma_j \min_s [E(i_r, j_s) - E(i_t, j_s)] \leq \varepsilon_{cut}, i \neq j \tag{2.2}$$

Unfavorable orientations are removed from the set of isomers that are compatible with a low-energy structure ε_{cut} from the conformation that is energetically best. By pruning the possibilities with DEE, the A* search algorithm can then evaluate all remaining possibilities and find all combinations of side chains that yield lowest-energy structures using a best-first search; alternative conformations within a designated energy cutoff may also be found to provide an ensemble of representative states for a target sequence. Unlike stochastic approaches, DEE/A* is deterministic and able to converge to a solution every time, though this may happen very slowly if few or no rotamers can ever be discarded. Methods like self-consistent mean field approximations are also available and used in protein design, and can often converge more quickly to a solution that DEE/A*; exceptions do exist in which convergence can be comparatively slower or even less accurate, at times, but this largely depends on specific features of different protein regions.[79] Despite these differences in convergence time, there is also disparity between converged solutions: unlike self-consistent mean field approaches, DEE/A* can guarantee the global minimum energy solution every time for a given set of structural and energetic constraints.

Backbone representations can also influence the accuracy of calculations in computational protein design methods. Efforts have been made towards including flexible backbone design into the energetic calculations,[93, 94, 95, 96, 97, 98] in which perturbations in ϕ- and ψ-angles are also allowed; this will naturally increase the amount of computation required for a given sequence. A reasonable alternative is to have multiple backbone representations for a given protein, either derived from simulation or by using multiple crystal structures. Regardless of the approach, these considerations provide a better idea of how the protein ensemble should behave, rather than just a single state.

Application to heterotrimeric G-proteins. DEE/A* can be used with any protein, as long as a backbone structure is provided, and we have selected a heterotrimeric G-protein (PDB: 1GP2, $G_i\alpha_1\beta_1\gamma_2$) as a model system.[99] Heterotrimeric G-proteins are important in many signal transduction pathways, and this protein family is comprised of 21 α-, 5 β- and 6 γ- (one splice variant) subunits; *in vitro* studies have suggested a degree of binding specificity between different subunits may restrict protein complex formation so that only specific combinations can be biologically active. [19, 11] A complete heterotrimer is activated after a ligand binds to a G-protein coupled receptor, which induces a conformational change in the α-subunit. This allows an exchange of GDP for GTP due to naturally higher concentrations of GTP in the cell, and results in an additional conformational change that favors a separation between the α-subunit from the $\beta\gamma$-heterodimer. Here, DEE/A* is used to systematically evaluate how all possible single mutants affect protein fitness in terms of structural stability and binding interactions. Mutational profiles can be constructed to describe how tolerant each wild-type side chain may be to amino-acid substitution, and the significance of each in maintaining stabilizing or binding interactions can be distinguished.

2.2 Methods

Wild-type conformations of the G-protein heterotrimer were taken from a 350-ns molecular dynamics simulation, by taking evenly spaced intervals 50-ns apart;[100] five consecutive snapshots were taken from each interval for a total of 40 unique backbone conformations. Amino-acid orientations were defined using an augmented version of the original Dunbrack–Karplus rotamer library:[101] $\pm 10^\circ$ were added to each χ_1- and χ_2-angle for finer, but still tractable, sampling. To discard unfavorable amino acids earlier on, a flexible rotamer approach was used to evaluate the possibilities, by grouping side-chain orientations with similar χ_1- and χ_2-angles.[102] Mutations were introduced to each position in the wild-type sequence, and wild-type side chains within 5 Å of the mutated position could re-orient itself

to a more favorable conformation if one were available—otherwise, all side chains and the protein backbone remained fixed. (See **Fig. 2.1**)

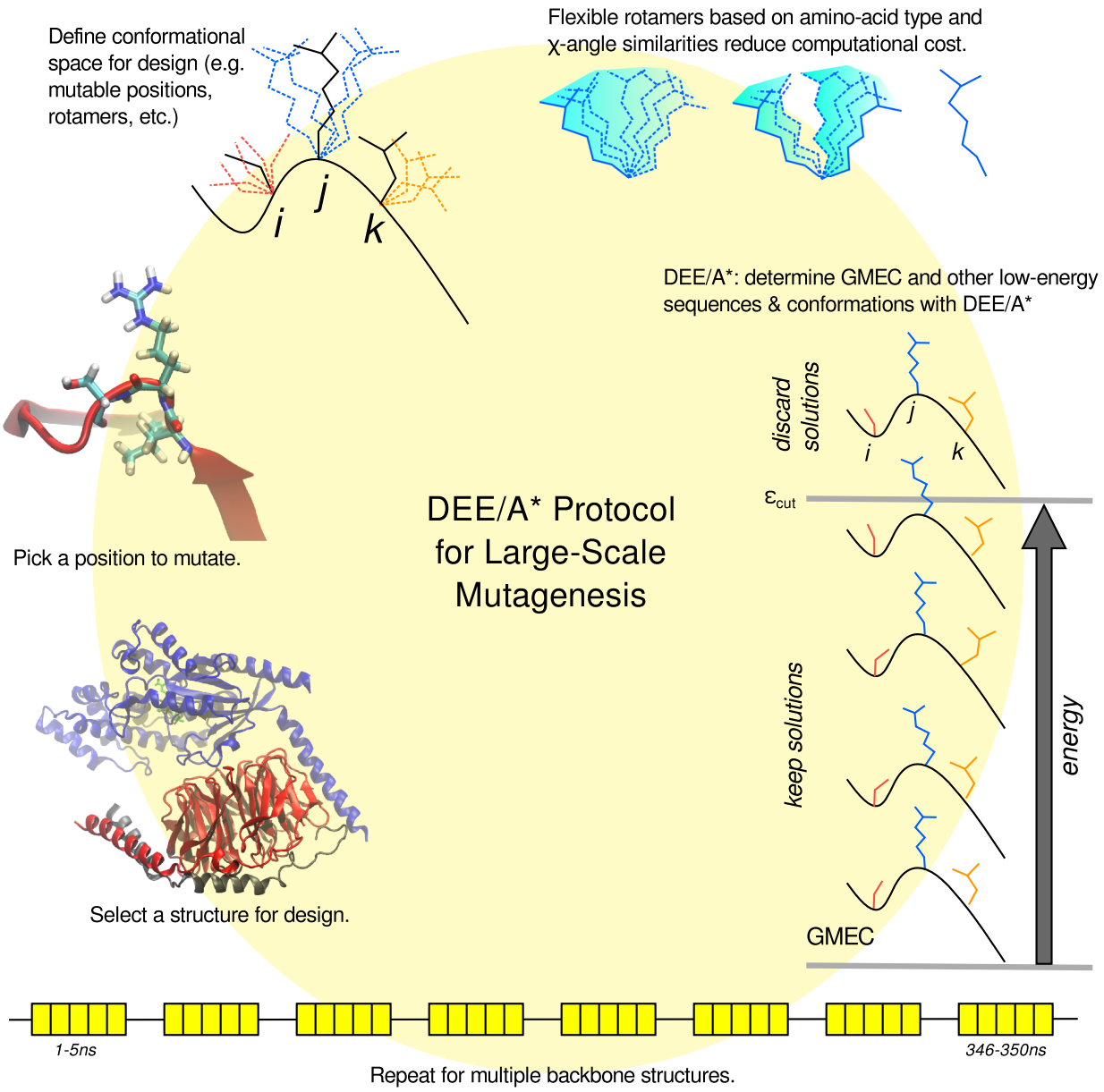


Figure 2.1: A wild-type conformation is selected from a molecular dynamics simulation, then a position is chosen for mutation. Amino acids and their rotamers are defined by a rotamer library, and evaluated as possible substitutions to the wild-type side-chain conformations. Flexible rotamers are defined by based on how similar their χ_1- and χ_2-angles are. For a given sequence, the best conformation is determined, then additional sequences and corresponding structures are also found within a given energetic cutoff from the best solution. This process can be repeated for any number of positions and wild-type conformations.

Hierarchy of energetic models. Energy calculations can be performed using any force field, and a variety of energetic models can be implemented in a hierarchical order, with the most computationally intense being applied last to approximate and discard poor solutions sooner. This strategy reduces the number of structures for later evaluation in the final steps,[103] in a way that utilizes computational resources more effectively. Protein conformations were initially evaluated using a distance-dependent dielectric ($\epsilon = 4r$),[61] and low-energy sequences were defined as being within 30 kcal/mol from the global minimum energy conformation. After preliminary pruning, the remaining solutions were assessed further using the generalized-Born with switching package (GBSW) in CHARMM.[104]

Evaluating Gα–GDP interactions. Gα wild-type side chains within 5 Å of any GDP atom in each of the 40 wild-type starting conformations were included in the analysis of Gα–GDP interactions. Not all side chains interact consistently with the ligand throughout the simulation due to backbone fluctuations, and the free energy data associated with these sequences were normalized accordingly.

Computational resources. All DEE/A* mutations for an individual mutation were performed on a single 3.4 Ghz Intel Pentium IV Xeon processor; most positions required about 4-5 hours of computing time. There are 685 mutable positions in 1GP2, and using a cluster with 235 processing nodes, an average of 48 hours are required to perform mutagenesis at all positions. Mutation free energy calculations required about 30 minutes of computing time on the same cluster.

Effective temperature for energetic calculations. A Boltzmann-weighted average was computed for each sequence to determine energetic differences over all 40 protein conformations. Effective temperatures were tested for a few systems of hydrogen bonds as a way to calibrate how temperature variation would influence the computed energy (**Fig. 2.2**). Temperatures were tested for 300K , 1500K, 3000K, 4500K, 6000K, 9000K, and 300,000K, to evaluate how much the Boltzmann distribution would shift, and how close it would approach the expected energy of a typical hydrogen bond (**Tables 2.1**, **2.2**, and **2.3**). Most hydrogen

bonds have about 3–7 kcal/mol of energy, depending on the pair of hydrogen bond donors and acceptors involved; for the mutations in the three systems of hydrogen bonds that were tested, a temperature of 4500K yields values closest to this range and was chosen as the effective temperature for all DEE/A* calculations.

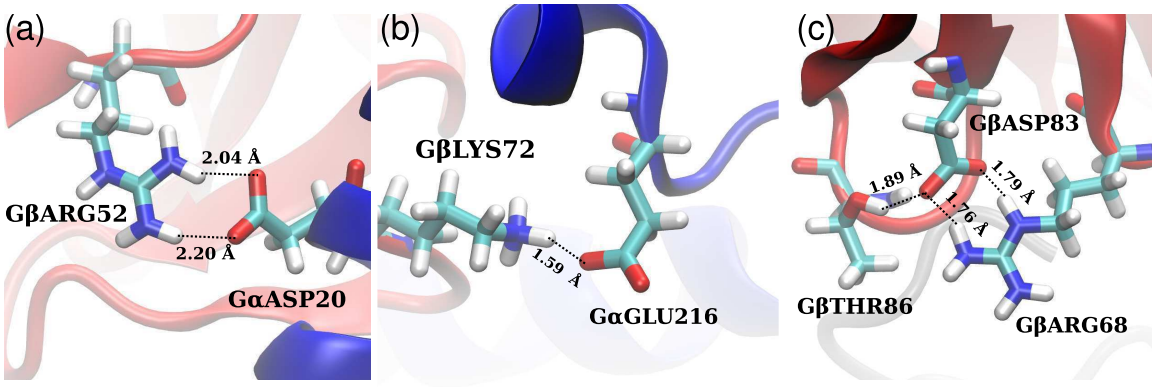


Figure 2.2: Effective energy was defined based on well-established intermolecular interactions and how well different temperatures could capture expected energetic changes due to amino-acid substitution in: (a) a doubly-bonded salt bridge, GαD20–GβR52, b) a singly-bonded salt bridge, GαE216–GβK57, and (c) a hydrogen-bond network between GβR68–GβE83–GβT86.

Table 2.1: $\langle \Delta\Delta G_{bind} \rangle$ in kcal/mol for select mutations at GαD20–GβR52 salt bridge.

Temp. (K)	GαD20A	GαD20E	GαD20N	GαD20Q	GβR52A	GβR52K
300	-1.5± 0.2	-13.1 ± 2.0	-4.8 ± 0.7	-7.2 ± 0.9	5.2 ± 1.0	0.0 ± 0.0
1500	-0.2 ± 0.1	-10.4 ± 1.0	-2.9 ± 0.4	-5.6 ± 0.5	1.8 ± 0.4	1.3 ± 0.1
3000	2.1 ± 0.1	-6.7 ± 0.5	0.4 ± 0.2	-2.6 ± 0.3	2.2 ± 0.2	1.8 ± 0.1
4500	3.3 ± 0.1	-4.6 ± 0.3	1.9 ± 0.2	-1.1 ± 0.2	2.4 ± 0.1	2.0 ± 0.1
6000	4.0 ± 0.2	-3.5 ± 0.3	2.7 ± 0.2	-0.3 ± 0.2	2.6 ± 0.1	2.2 ± 0.1
9000	4.9 ± 0.2	-2.4 ± 0.2	3.6 ± 0.2	0.7 ± 0.2	2.7 ± 0.1	2.5 ± 0.1
300,000	7.7 ± 0.3	0.3 ± 0.3	6.3 ± 0.3	3.3 ± 0.3	3.0 ± 0.1	3.2 ± 0.1

Table 2.2: $\langle \Delta\Delta G_{bind} \rangle$ in kcal/mol for select mutations at GαE216–GβK57 salt bridge.

Temp. (K)	Gα E216A	Gα E216D	Gα E216N	Gα E216Q	Gβ K57A	Gβ K57R
300	2.0 ± 0.3	-2.6 ± 0.4	-8.7 ± 1.4	-7.6 ± 1.1	-8.0 ± 1.3	-10.2 ± 1.6
1500	1.2 ± 0.3	-0.7 ± 0.1	-4.5 ± 0.7	-7.0 ± 0.5	-8.0 ± 1.3	-10.2 ± 1.6
3000	1.1 ± 0.2	1.3 ± 0.1	-1.5 ± 0.3	-5.0 ± 0.2	-7.9 ± 1.3	-10.0 ± 1.6
4500	1.9 ± 0.1	2.3 ± 0.1	-0.2 ± 0.2	-4.2 ± 0.2	-6.4 ± 1.1	-7.8 ± 1.3
6000	2.4 ± 0.1	2.8 ± 0.1	0.5 ± 0.2	-3.8 ± 0.1	-3.1 ± 0.8	-4.4 ± 0.9
9000	2.9 ± 0.1	3.3 ± 0.1	1.2 ± 0.1	-3.3 ± 0.1	2.3 ± 0.4	-0.3 ± 0.4
300,000	3.9 ± 0.1	4.4 ± 0.1	2.4 ± 0.1	-2.5 ± 0.1	8.0 ± 0.1	4.6 ± 0.2

Table 2.3: $\langle \Delta\Delta G_{bind} \rangle$ in kcal/mol for select mutations in the hydrogen-bond network

Temp. (K)	Gβ R68A	Gβ R68K		
300	9.3 ± 1.3	-2.9 ± 0.4		
1500	3.7 ± 0.4	-1.4 ± 0.3		
3000	3.1 ± 0.2	-0.1 ± 0.2		
4500	3.2 ± 0.2	0.6 ± 0.1		
6000	3.3 ± 0.2	1.0 ± 0.1		
9000	3.4 ± 0.2	1.4 ± 0.1		
300,000	3.8 ± 0.1	2.3 ± 0.1		
	Gβ D83A	Gβ D83E	Gβ D83N	Gβ D83Q
300	-5.3 ± 0.9	-1.5 ± 0.2	-4.3 ± 0.7	-5.3 ± 0.8
1500	-4.3 ± 0.6	2.8 ± 0.3	-3.2 ± 0.4	-3.1 ± 0.6
3000	-1.8 ± 0.3	4.2 ± 0.2	-1.4 ± 0.3	-1.0 ± 0.3
4500	-0.1 ± 0.2	4.8 ± 0.2	0.0 ± 0.2	0.3 ± 0.2
6000	0.8 ± 0.2	5.1 ± 0.2	0.9 ± 0.1	1.0 ± 0.2
9000	1.8 ± 0.1	5.7 ± 0.2	1.9 ± 0.1	1.9 ± 0.2
300,000	3.5 ± 0.2	7.1 ± 0.2	4.0 ± 0.2	3.8 ± 0.2
	Gβ T86A	Gβ T86S	Gβ T86C	
300	-3.1 ± 0.5	-2.1 ± 0.3	-3.0 ± 0.3	
1500	-1.1 ± 0.2	-0.5 ± 0.1	-2.2 ± 0.2	
3000	0.3 ± 0.1	0.1 ± 0.1	-0.9 ± 0.1	
4500	0.9 ± 0.1	0.3 ± 0.1	-0.3 ± 0.1	
6000	1.2 ± 0.1	0.5 ± 0.1	0.0 ± 0.1	
9000	1.5 ± 0.1	0.8 ± 0.1	0.4 ± 0.1	
300,000	2.5 ± 0.1	1.9 ± 0.1	1.4 ± 0.1	

Energy cutoff for defining fitness. Mutations for any protein sequence can be classified as being more favorable, worse or essentially similar to the wild-type sequence. Neutral mutations are expected to represent a significant proportion of all possible sequences,[105, 106] and the energy relative to wild-type should only have modest differences; by taking neutral sequences into consideration, a boundary between favorable and unfavorable states needs to be drawn. An appropriate definition for this cutoff was determined by evaluating the proportion of sequences remaining when sequences that simultaneously satisfy stability and binding interactions are removed at different cutoffs (**Figs. 2.3** & **2.4**). A cutoff at 2 kcal/mol would suggest that about 20% of the DEE/A* sequences should behave like wild-type, but this may be an overestimation; a more conservative cutoff of 1.5 kcal/mol was chosen instead.

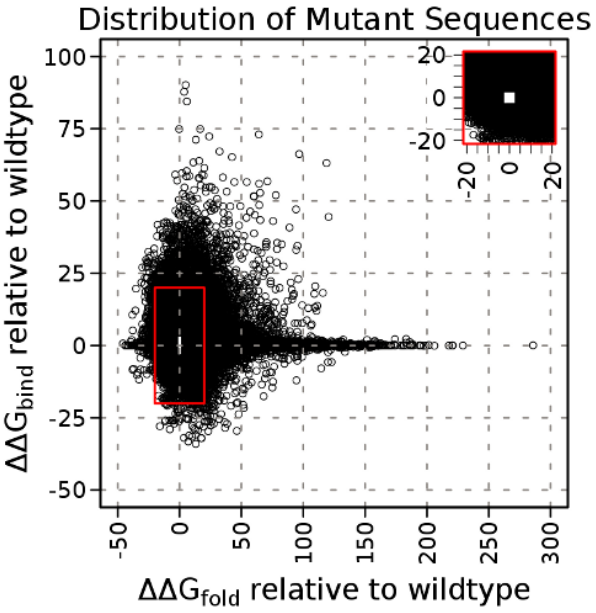


Figure 2.3: Sequences within ±1.5 kcal/mol are left out of the energy landscape to illustrate how different energy cutoffs were tested in finding a suitable energy cutoff for fitness.

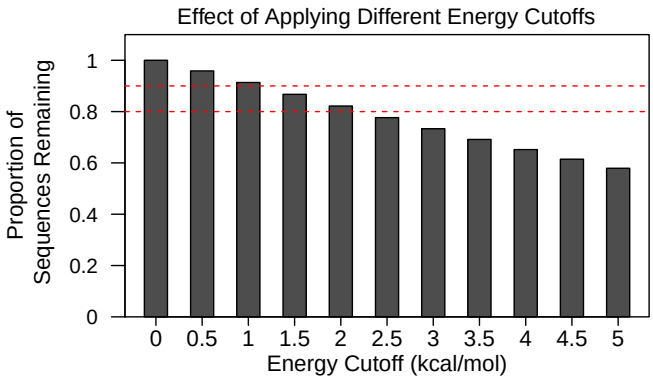


Figure 2.4: Energetic cutoffs were tested at 0.5-kcal/mol intervals, up to 5 kcal/mol. The proportion of remaining sequences after removing values within the designated energy cutoff is shown. Red lines are shown as a visual guide for 0.8 and 0.9 of proportion of unremoved data.

2.3 Results

2.3.1 Completeness of conformational sampling

Sufficiency of sample size. Including additional backbone conformations for performing DEE/A* is expected to improve the accuracy of our sampling, and these data can show how mutations behave consistently over an ensemble of structures. To determine whether or not a sufficient number of representations are used in our analysis, we considered the average energy difference between sequential subsets of intervals. Only the energy data for stability were considered in this analysis, since variance in binding energy is naturally low due to a small number of positions actually involved in binding. From our data, we defined the following subsets of conformations (**Table 2.4**):

Variance in energy measuring structural stability tends to have more variation relative to binding interactions, because most, if not all, positions must make some contribution towards it; very few positions are often found at binding interfaces, while a properly folded and stable protein is a pre-requisite for correct protein function. Consequently, this analysis on how energetic averages changed with an increasing number of protein conformations in

Table 2.4: Indexed protein structures are listed in intervals according to nanosecond in simulation.

A: [1,5]
B: [1,5] ∪ [48,52]
C: [1,5] ∪ [48,52] ∪ [98,102]
D: [1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152]
E: [1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202]
F: [1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252]
G: [1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252] ∪ [298, 302]
H: [1,5] ∪ [48,52] ∪ [98,102] ∪ [148,152] ∪ [198, 202] ∪ [248, 252] ∪ [298, 302] ∪ [346, 350]

sampling focused only on stability, $\Delta\Delta G_{fold}$. The average energy of each mutant sequence was computed for each subset, and we measured how this average energy will change from one subset to another, in sequential order, by taking the absolute difference between corresponding sequences: $|\langle A\rangle - \langle B\rangle|$, $|\langle B\rangle - \langle C\rangle|$, $|\langle C\rangle - \langle D\rangle|$, and so forth. The difference in energy variation as the number of structural ensembles increased was partitioned into 1-kcal/mol bins, and the distribution of variance across them was measured to summarize how sampling improves the consistency of our energetic data (**Fig. 2.8**). By including additional snapshots, the number of outliers in our protein sequence space could be reduced dramatically. In practice, eliminating outliers entirely might not be possible, if flexible regions exist in the protein being studied; it is expected that highly flexible proteins will require more structural conformations for analysis. Furthermore, it appears that the proportion of sequences with an average difference greater than 2 kcal/mol becomes less apparent when 20 or more conformations are used in our calculations, and thus having a total of 40 unique protein conformations provides adequate sampling for this protein.

Selection of rotamer library. We were interested in seeing how well the augmented Dunbrack–Karplus library ($\pm 10^\circ$ to each χ_1- and χ_2-angle) performed by taking DEE/A* results from a single backbone conformation, and applying a Newton–Rhapson energy minimization algorithm to it (**Fig. 2.5**). This analysis only focuses on changes in $\Delta\Delta G_{fold}$, because very few positions have any variation in binding interactions overall. Approximately 14% of the sequences were found to be unfavorable in one approach and favorable in the

other. Unfavorable states remained unfavorable in about 60% of the sequences, while favorable sequences remained favorable in about 20% of the data; about 7% of over 6000 single-mutant sequences could be improved in a meaningful way using energy minimization. Discrepancies in energy calculations tend to arise with the aromatic amino acids, or the charged ones (**Fig. 2.6**). Most of the energetic improvements that arise from off-rotamer sampling are very modest: the majority of energy differences between the two methods are within 5 kcal/mol of each other, prior to any adjustment with effective temperature (**Fig. 2.7**). Furthermore, minimization of the wild-type structure also contributes to these energetic discrepancies between the two methods of calculation, by lowering the energy of the reference state.

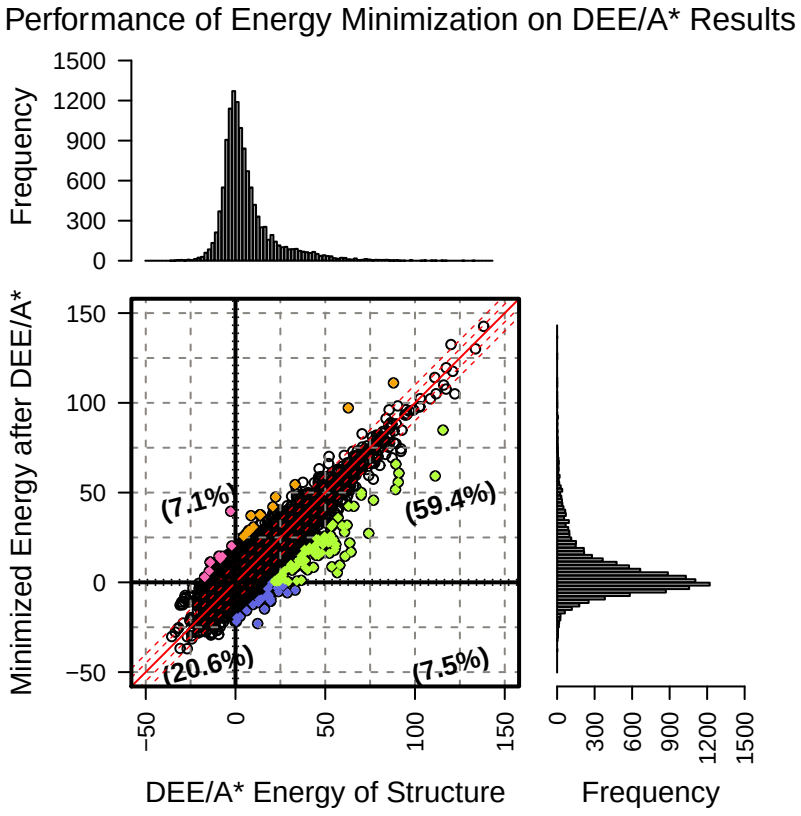


Figure 2.5: All DEE/A* solutions from one snapshot were minimized using a Newton–Rhapson algorithm, and the results were compared to the original DEE/A* energies. The proportion of data found in each quadrant is labeled. Red line indicate the 1-1 correlation (solid) and then -10, -5, 5 and 10 kcal/mol from this best-fit line as a visual reference. Histograms on top and to the side show the density of how these structures are distributed, according to the two approaches for computing energy.

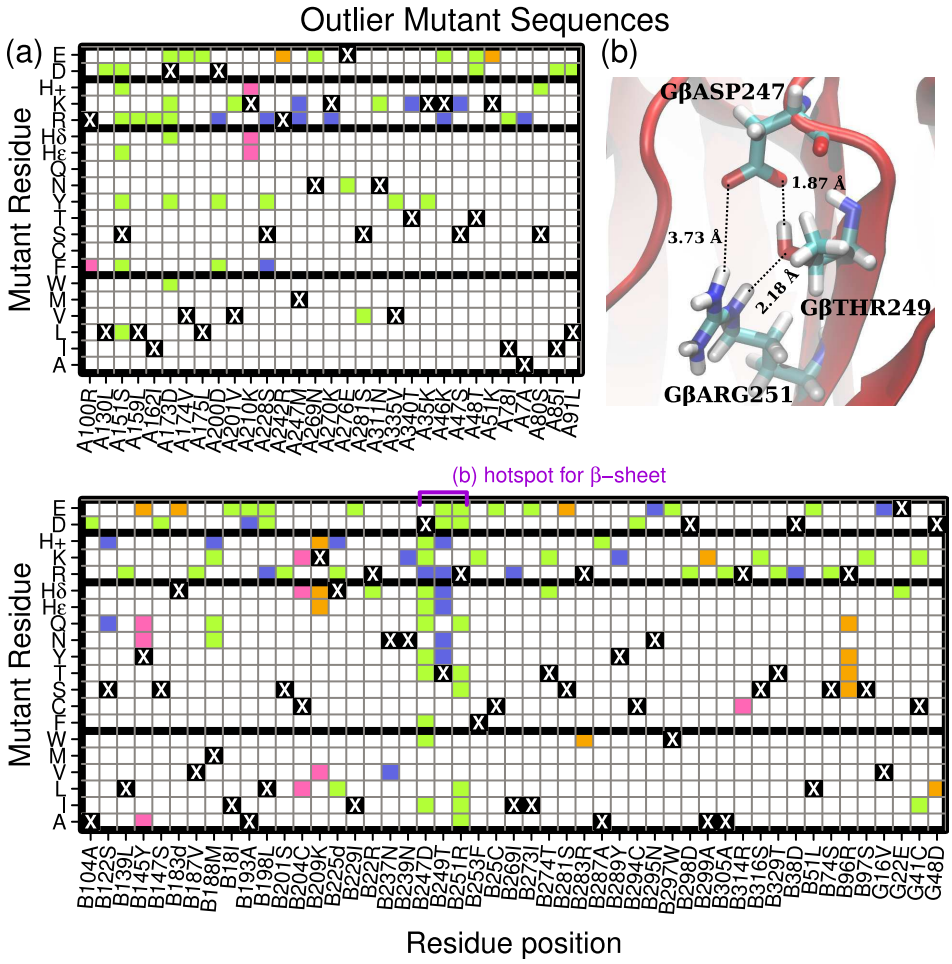


Figure 2.6: Sequences in the colored regions of **Figure 2.5** are shown here, organized according to position and mutation. Colors continue to correspond with the regions labeled in **Figure 2.5**. In (b), a structural example on Gβ is shown for the hot spot of sequences found to be very different after minimization.

2.3.2 Mutational robustness from computational mutagenesis

Mutagenesis profiles from implementing DEE/A*. Many mutations have a neutral effect on the protein, but there is a tendency for mutations to be less favorable than wild type. Approximately two-thirds of the sequences explored by DEE/A* are destabilizing to the wild-type structure, and greater energetic variance is seen in these sequences than those measured for changes in binding interactions (**Fig. 2.9**). This is due both to having a smaller number of amino acids overall involved in binding, and to having a broad range

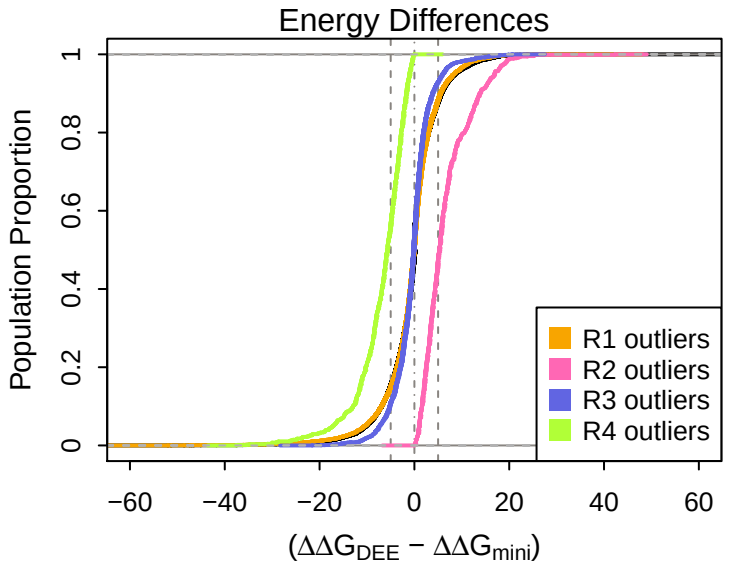


Figure 2.7: The differences found from computing energy using the two methods were calculated, and shown here as a cumulative distribution function. Regions correspond to **Figure 2.5** quadrants, and colored accordingly.

of microenvironments—from hydrophobic to highly solvent exposed—available in the folded protein. A complete sequence profile for every position was established for our model system, identifying specific regions of unfavorable amino-acid substitution and highlighting those that are less sensitive to mutation (when both stability and binding are considered, **Fig. 2.10** & **2.11**, how changes are made to stability only, **Fig. 2.12** & **2.13** and also how binding interactions are affected, **Fig. 2.14** & **2.15**). Positions with several allowable and favorable substitutions are usually a direct consequence of having fewer geometric or electrostatic constraints; when very diverse functional groups cannot be accommodated at a position, it suggests that unique side-chain interactions exist in the region and are required to maintain protein fitness.

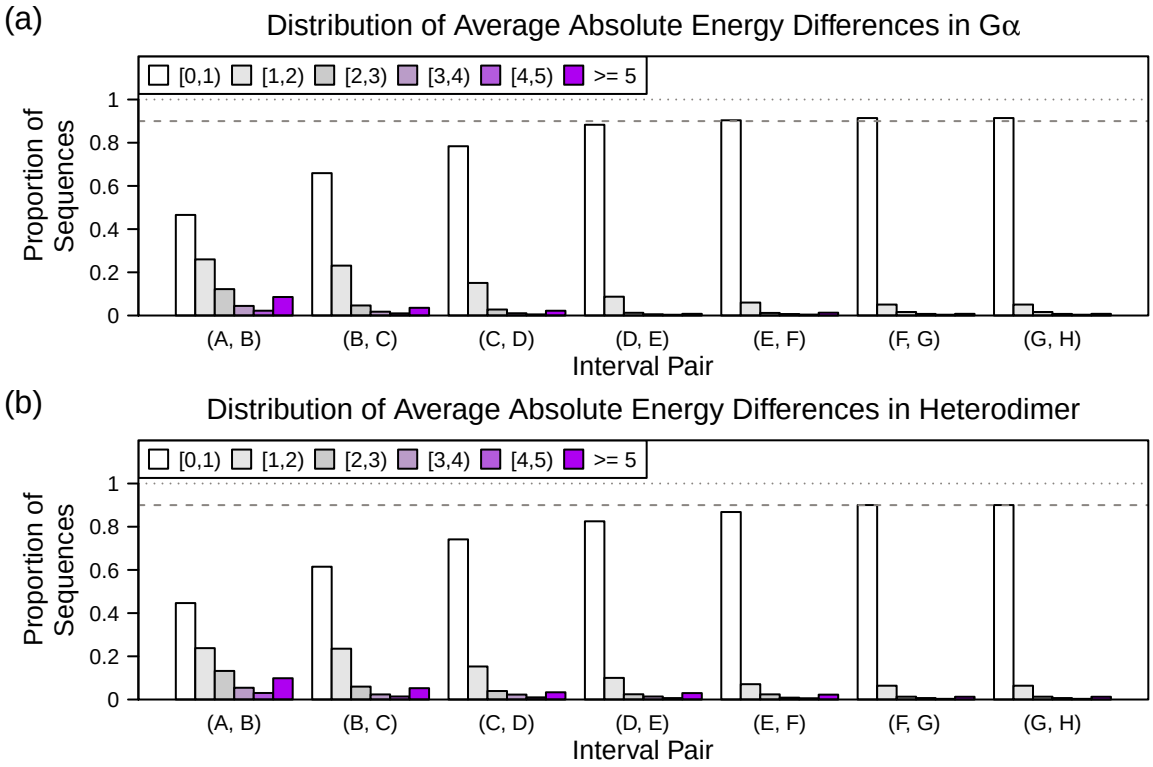


Figure 2.8: For each position, $\langle \Delta\Delta G_{fold} \rangle$ is computed for different subsets of conformations used in the data set (see **Table 2.4**.) The proportion of data at [0,1)-, [1,2)-, [2,3)-, [3,4)-, [4,5)-kcal/mol or ≥ 5 kcal/mol as the number of conformations included in overall sampling increases. Sequences for Gα and the $\beta\gamma$-heterodimer show similar patterns in convergence, as the ensemble of states used to represent a sequence increases.

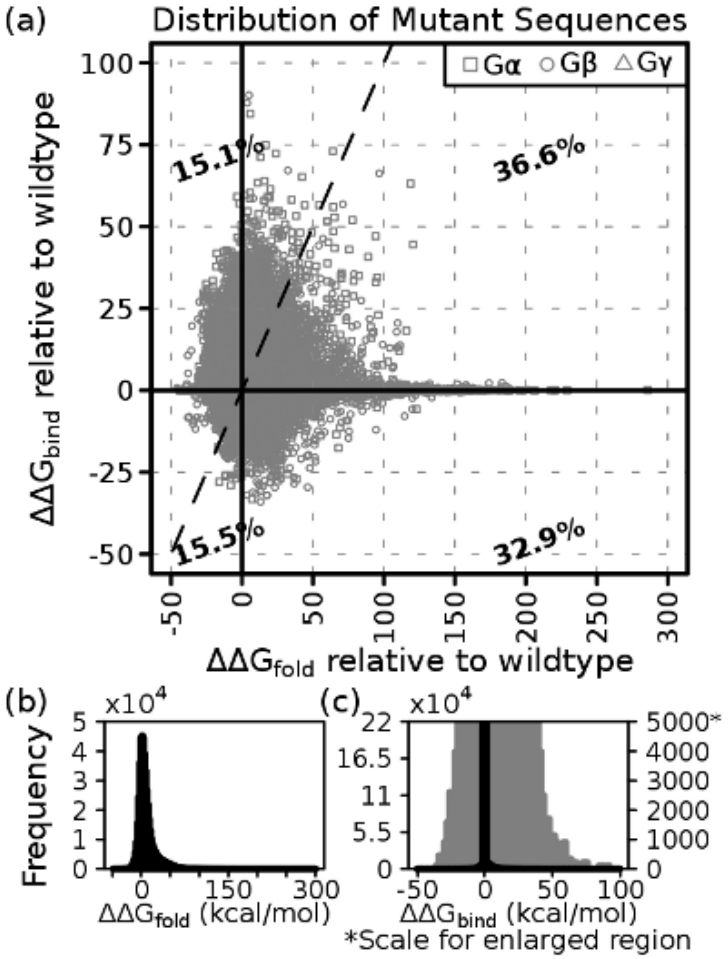


Figure 2.9: All mutant sequences resulting from DEE/A* are shown, and are referenced to the wild-type sequence. (a) The distribution of each sequence is mapped change in stability ($\Delta\Delta G_{fold}$) and change in binding ($\Delta\Delta G_{bind}$). The density of the distribution is shown in panels (b) and (c) for each, $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$, respectively. Mutations are distributed asymmetrically for $\Delta\Delta G_{fold}$; though also asymmetric, sequences are generally centered around 0 kcal/mol for $\Delta\Delta G_{bind}$, due to a high proportion of sequences uninvolved in binding.

Figure 2.10: Stability and binding are both important requirements, and thus $\max(\langle \Delta\Delta G_{fold} \rangle, \langle \Delta\Delta G_{bind} \rangle)$, the maximum energy of either stability or binding, is evaluated here for all mutations at every position. Unfavorable substitutions are shown in red, while neutral changes or favorable ones are shown in white and blue, respectively.

Figure 2.11: If stability or binding interaction requirements are not met, then the mutation will result in a loss of function. To evaluate this criteria, $\max(\langle \Delta\Delta G_{fold} \rangle \, , \langle \Delta\Delta G_{bind} \rangle)$, the maximum of either fitness aspect is considered for every position and all substitutions that are made to it. Red indicates an unfavorable change, while white and blue reflect neutral and favorable substitutions, respectively.

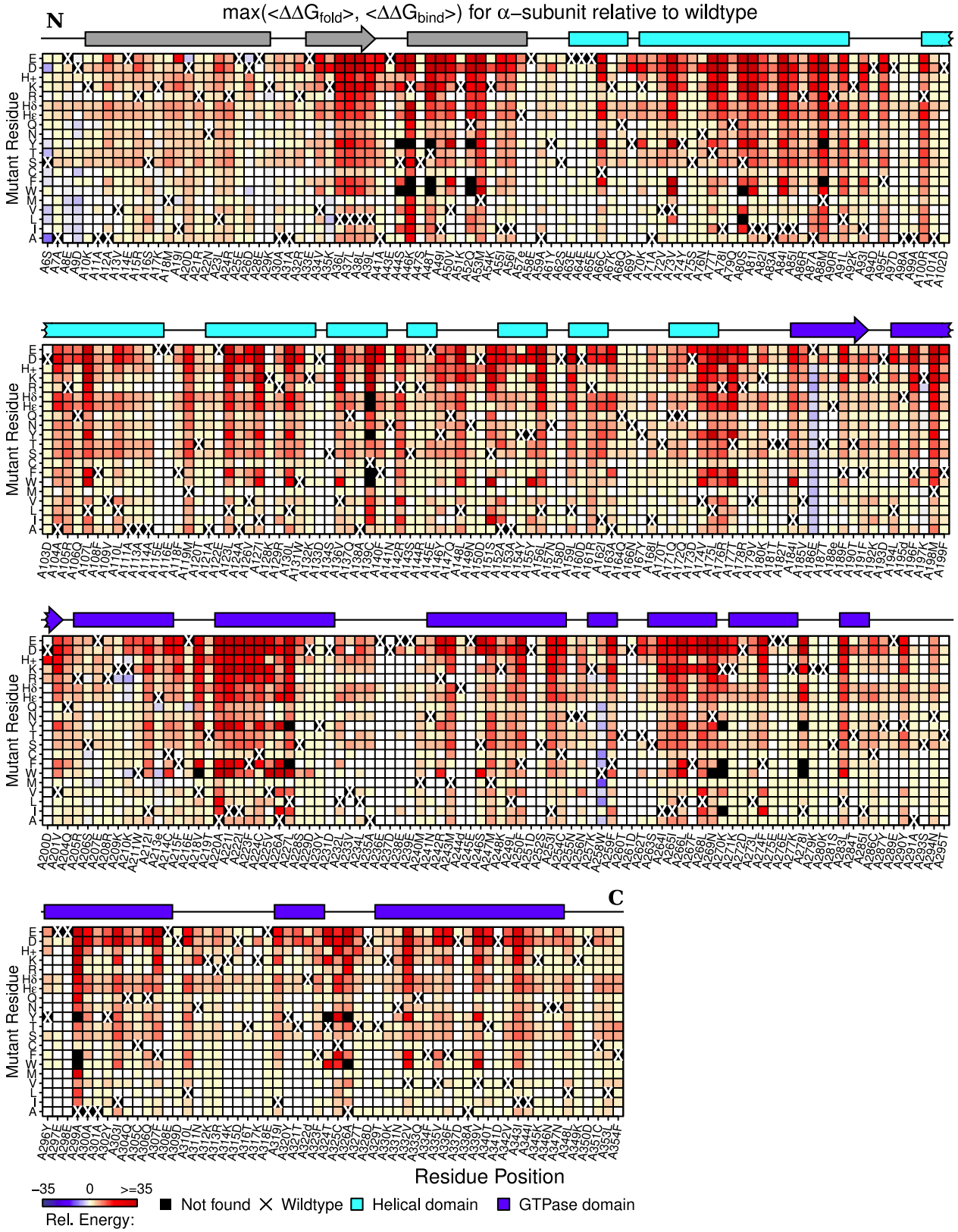

N
max(<ΔΔG_fold>, <ΔΔG_bind>) for α–subunit relative to wildtype
C
Mutant Residue
Residue Position
–35 0 >=35
Rel. Energy:
Not found
Wildtype
Helical domain
GTPase domain

Figure 2.10

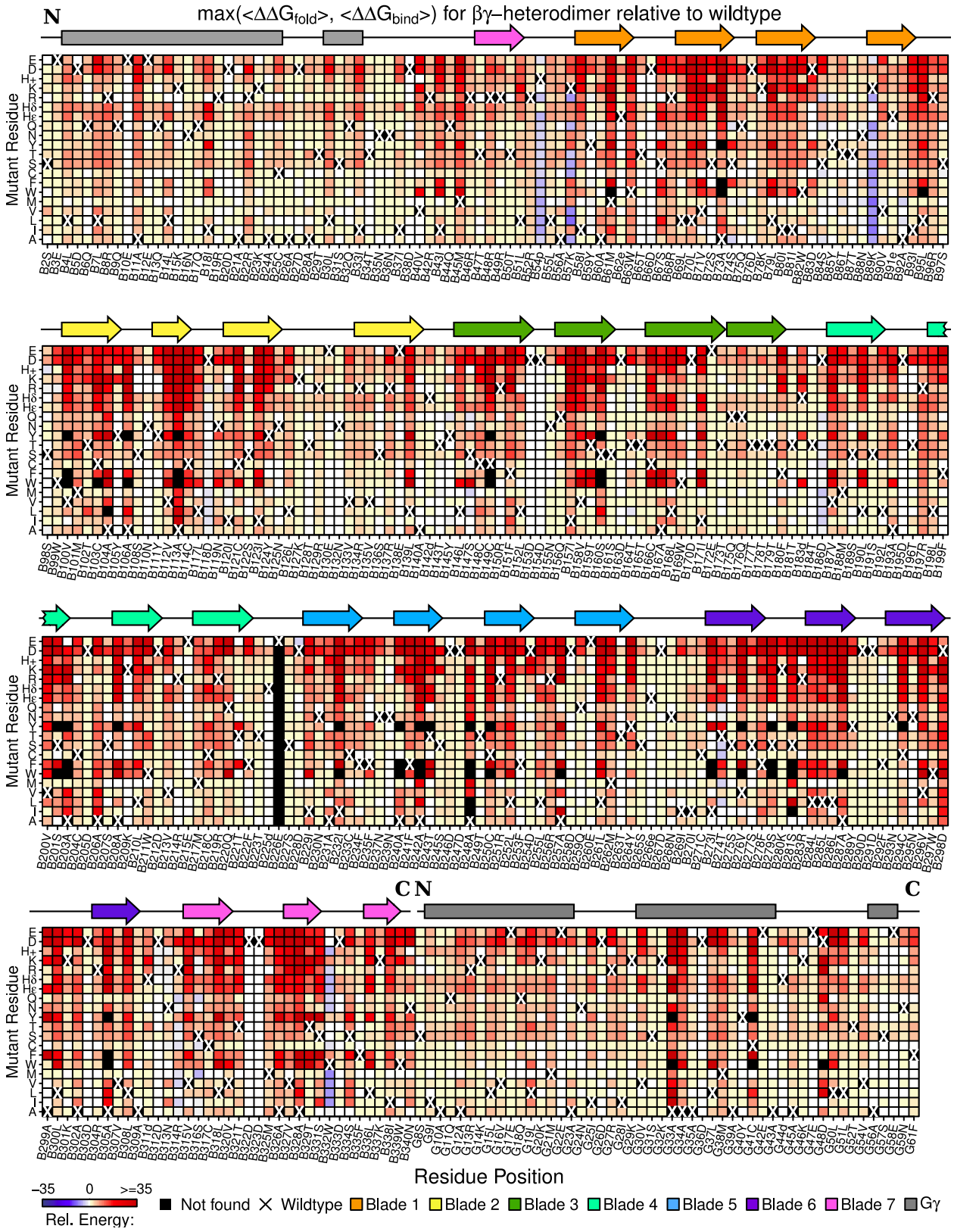

N
max(<ΔΔG_fold>, <ΔΔG_bind>) for βγ–heterodimer relative to wildtype
Mutant Residue
C N
C
Residue Position
–35 0 >=35
Rel. Energy:
Not found
Wildtype
Blade 1
Blade 2
Blade 3
Blade 4
Blade 5
Blade 6
Blade 7
Gγ

Figure 2.11

Figure 2.12: The average energy difference in stability of each mutant sequence, relative to the wild type, $\langle \Delta\Delta G_{fold} \rangle$, is mapped out for every position in the protein. Unfavorable substitutions are shown in red, while white and blue correspond to neutral and favorable mutations, respectively. Secondary structure for the α subunit is also shown, and the major domains are differentiated in color.

Figure 2.13: Changes in average energy with respect to the wild-type sequence is shown here for all mutations in the $\beta\gamma$-heterodimer, $\langle \Delta\Delta G_{fold} \rangle$. Mutations that worsen wild-type fitness are shown in red; neutral mutations and favorable ones are shown in white and blue, respectively. Protein secondary structure is included: different colors are used for each β-propeller blade, and the γ subunit is shown separately at the end of the profile.

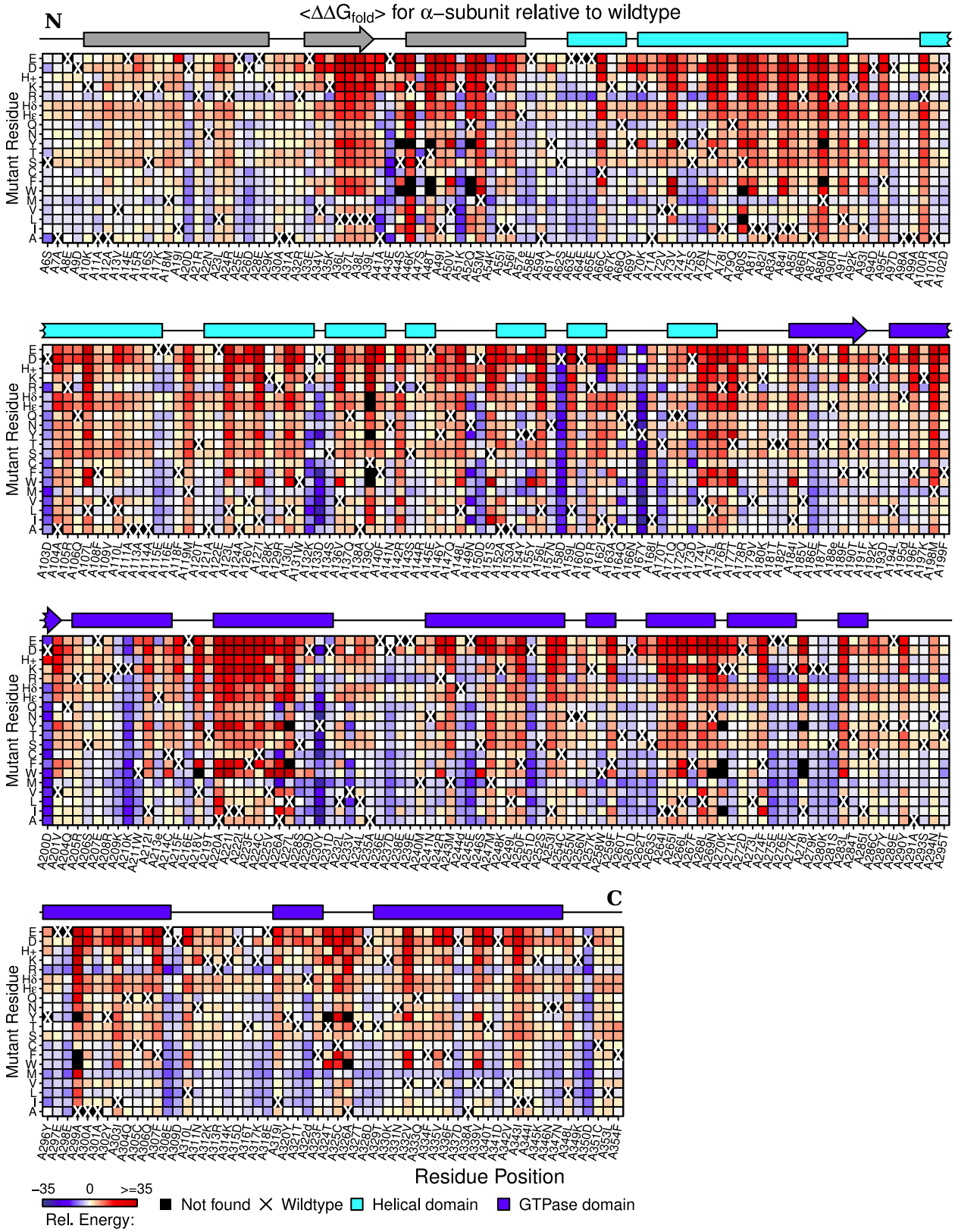

N
<ΔΔG_fold> for α−subunit relative to wildtype
Mutant Residue
C
Residue Position
−35
0
>=35
Rel. Energy:
Not found
Wildtype
Helical domain
GTPase domain

Figure 2.12

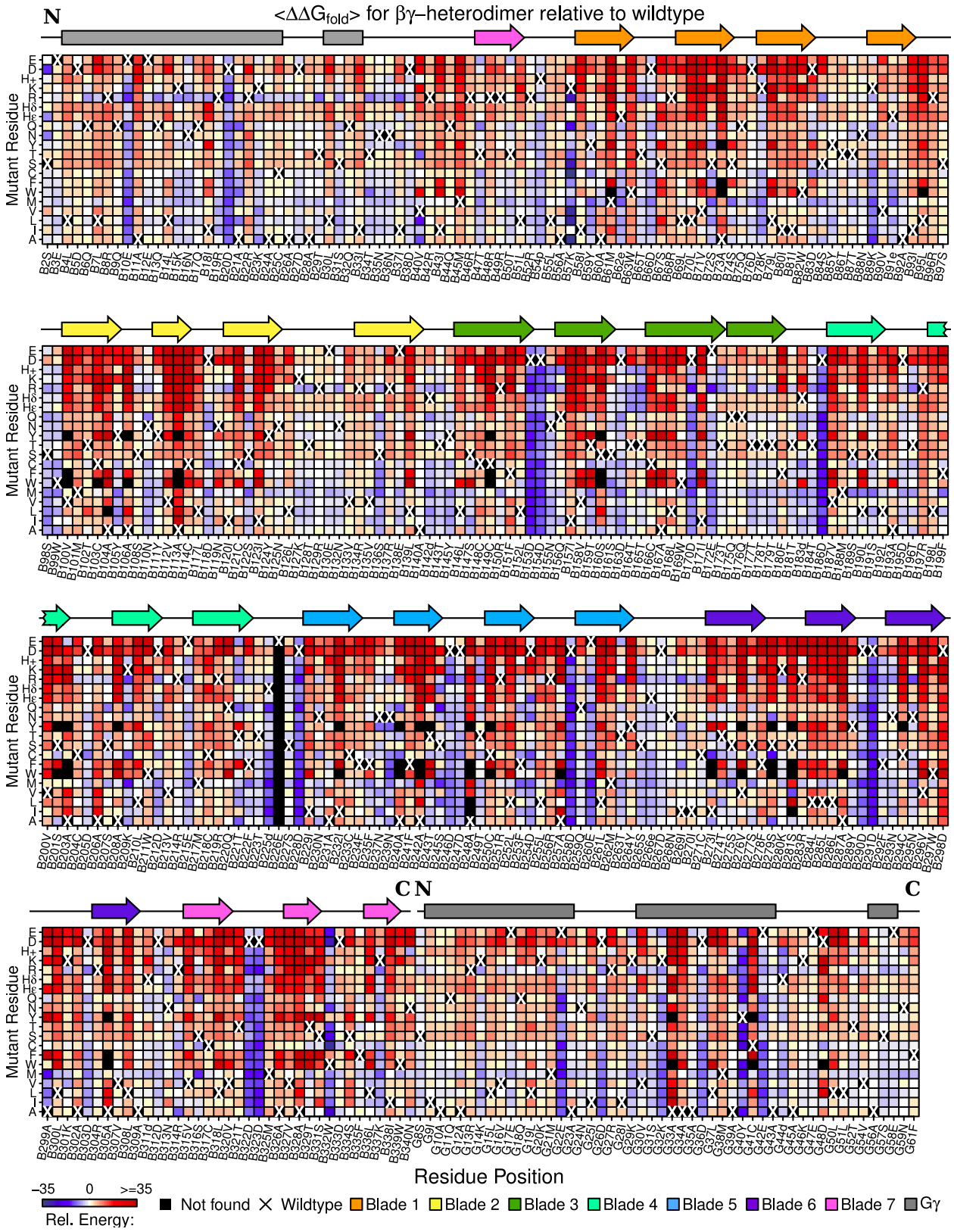

N
<ΔΔG_fold> for βγ–heterodimer relative to wildtype
Mutant Residue
C N
C
Residue Position
–35 0 >=35
Rel. Energy:
Not found
Wildtype
Blade 1
Blade 2
Blade 3
Blade 4
Blade 5
Blade 6
Blade 7
Gγ

Figure 2.13

Figure 2.14: Energetic changes in binding interactions, $\langle \Delta\Delta G_{bind} \rangle$, with the $\beta\gamma$-heterodimer is shown for all single mutants in the α subunit. Improvements relative to wild type is shown in blue, while mutations that worsen binding interactions are shown in red; neutral changes are in white. Protein secondary structure is indicated in this profile, and colored according to domain. Very few positions are at the Gα–$\beta\gamma$ binding interface, and so the majority of mutational effects will be neutral.

Figure 2.15: Mutational effects on binding interactions of the $\beta\gamma$-heterodimer with respect to the α subunit, $\langle \Delta\Delta G_{bind} \rangle$, is shown for all positions. Favorable, neutral and unfavorable changes are are colored blue, white and red, respectively. The domains in each part of the $\beta\gamma$-heterodimer is shown. Most mutations will have limited effects on overall fitness, since very few residues are at the binding interface.

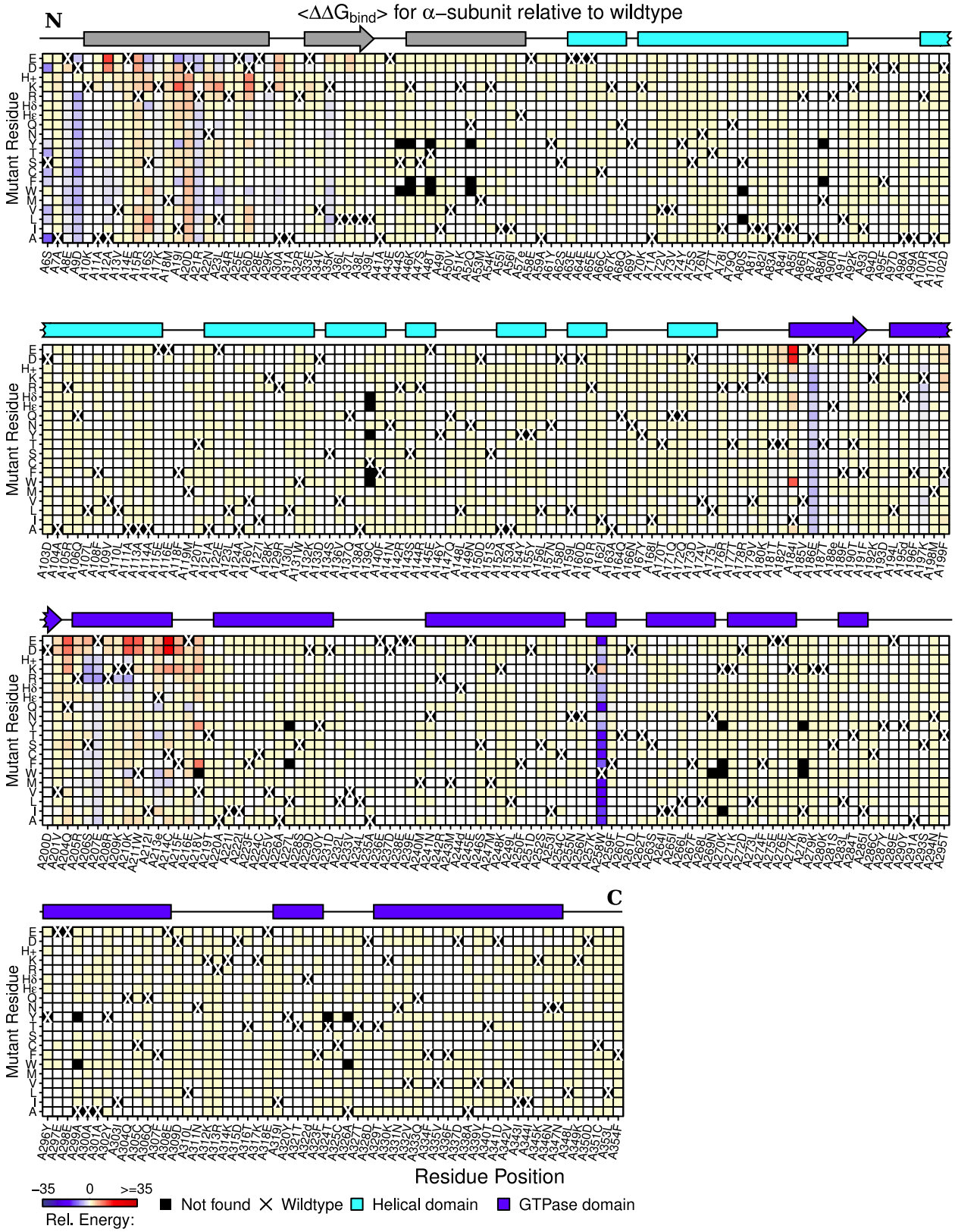

N
<ΔΔG_bind> for α−subunit relative to wildtype
Mutant Residue
C
Residue Position
−35
0
>=35
Rel. Energy:
Not found
Wildtype
Helical domain
GTPase domain

Figure 2.14

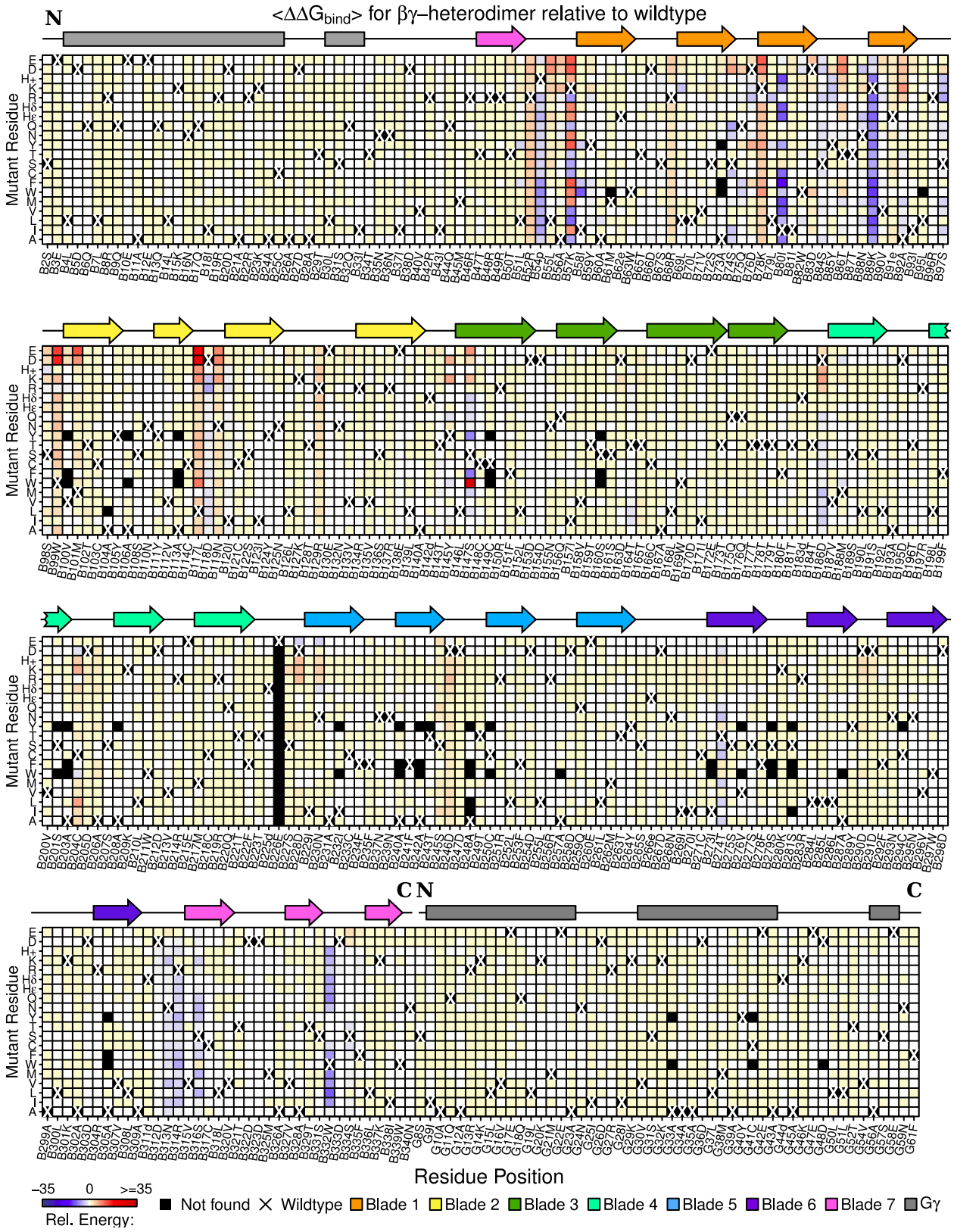

N
<ΔΔG_bind> for βγ–heterodimer relative to wildtype
Mutant Residue
C N
C
Residue Position
–35 0 >=35
Rel. Energy:
Not found
Wildtype
Blade 1
Blade 2
Blade 3
Blade 4
Blade 5
Blade 6
Blade 7
Gγ

Figure 2.15

Disentangling functional roles using energy profiles. Energetic profiles were created separately for stability and binding interactions using the free energy of all mutant sequences. (See Materials and Methods.) By measuring these two aspects of fitness independently, functional trade-offs in the protein can be identified, as demonstrated by the GDP-binding pocket of Gα. (**Fig. 2.16**) If either requirement for stability or binding is not satisfied, the overall fitness of the protein is worse than wild type—the maximum energy of either stability or binding, $\max(\Delta\Delta G_{fold}, \Delta\Delta G_{bind})$, can make this distinction for a given mutant. Asp150, for example, can be easily replaced by most amino acids and remain functional, because the native orientation points the carboxyl group away from GDP, despite being near a guanine nitrogen (**Fig. 2.16a**). Nearly all substitutions can improve protein stabilization at Ala41, Lys46, Lys270 and Lys180 in Gα, but the same mutations are poor candidates for binding GDP (**Fig. 2.16b**). Similarly, side-chain substitutions in the same subunit can improve binding interactions relative to wild type at positions Ile49, Asp200 and Asn149, but doing so will generally destabilize the α-subunit (**Fig. 2.16c**). Bulkier, aromatic amino acids do not fit well in this region, and charged side chains are also poor candidates because of their electrostatic requirements. These general trends are exhibited throughout the heterotrimer, and functional trade-offs are only a concern for the small number of positions present at the protein-binding interface or involved in protein–ligand interactions (**Fig. 2.14** & **2.15**). Most positions are sensitive to substitution, and this is expected for a highly evolved protein family. Gα positions at the amino terminus or in switch II (residues 202–209) have the greatest energetic variation after mutation than other positions in the subunit, and these regions are known to interact with the β-subunit when inactive.[107, 108, 2, 11] Gβ, an example of a WD40 β-propeller protein, has positions at the binding interface, that also show similar trend, and where stability is lost after mutation is consistent with what we expect from the WD40 protein family (**Fig. 2.17**).[11, 109]

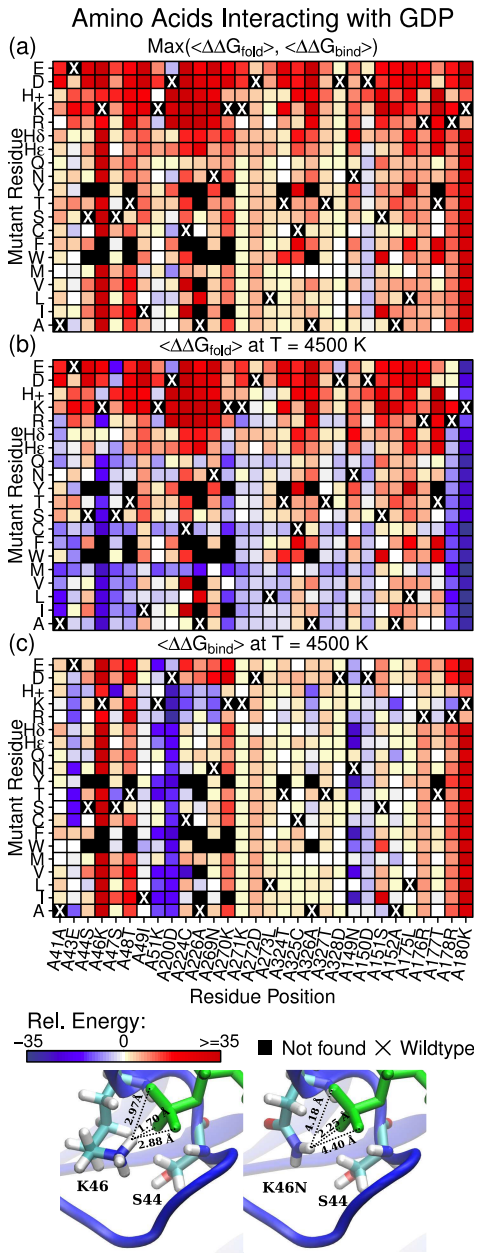


Figure 2.16: The number of positions in the Gα–GDP-binding pocket is relatively few, and are organized here according to whether they belong in the Ras-like GTPase or helical domain. Mutational effects are organized in different panels: (a) $\max(\langle\Delta\Delta G_{fold}\rangle, \langle\Delta\Delta G_{bind}\rangle)$, the maximum of either the worst mutation for stability or for binding, (b) $\langle\Delta\Delta G_{fold}\rangle$, the average change in stability after mutation and (c) $\langle\Delta\Delta G_{bind}\rangle$, the average change in binding interactions after mutation. Generally, most positions cannot be replaced, and thus panel (a) is mostly red (unfavorable energy, relative to wild type.) Functional trade-offs can be see any many positions: what is good for one aspect of fitness (blue), can be detrimental to the other (red); in panel (d), an example mutation K64N is shown to illustrate how a substitution can be unfavorable to binding, but a significant improvement in stability due to increased hydrogen bond interactions.

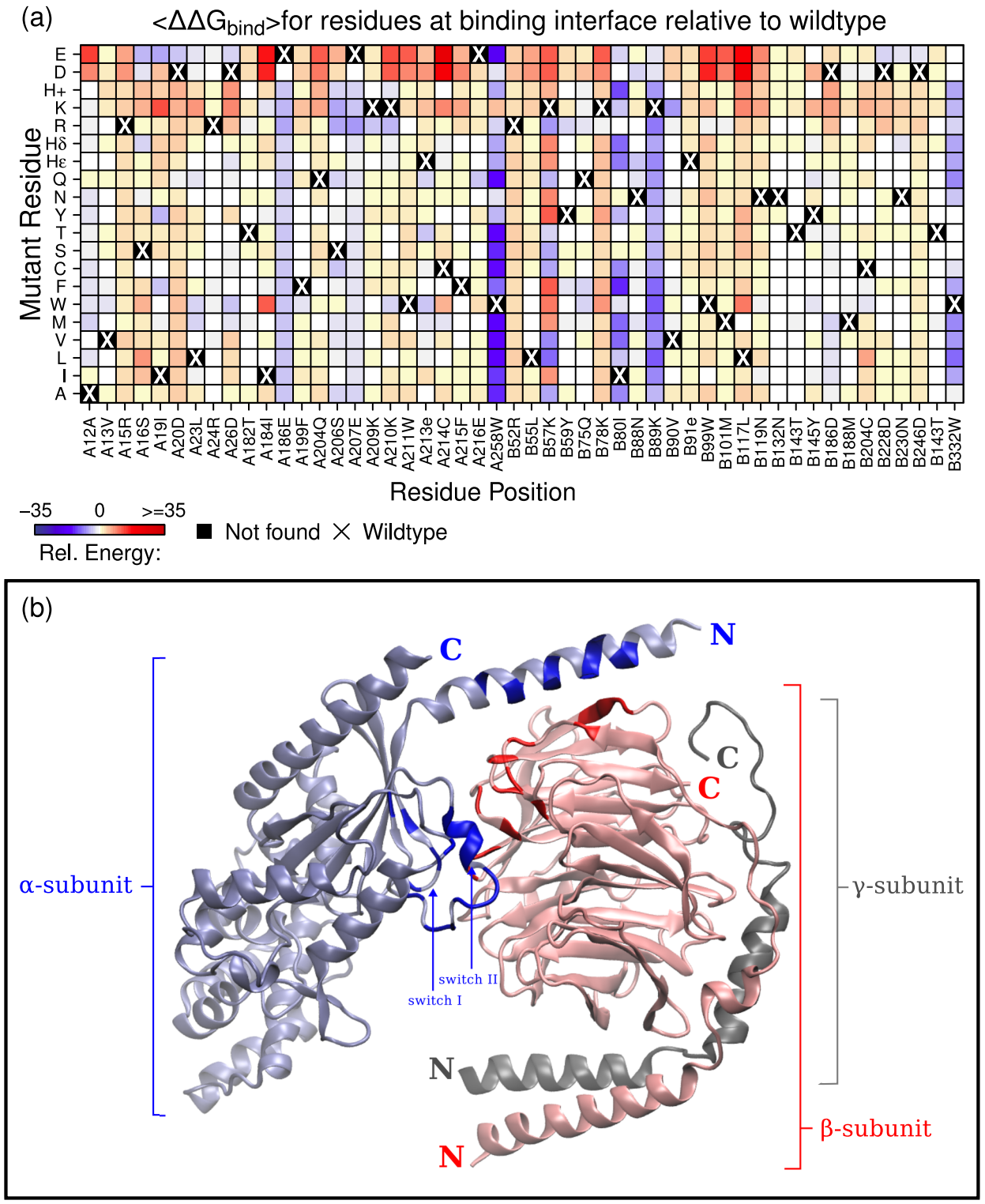


Figure 2.17: Positions that are sensitive to mutations and involved in protein–protein binding interactions are shown in panel (a). There are subsets of **Figure 2.14** & **2.15**, summarized for clarity. In (b), the positions that have any relevance in binding are mapped onto the structure, and generally reflect important regions of the G-protein heterotrimer that are known binding interfaces.

Figure 2.18: The average for each sequence, $\langle \Delta\Delta G_{fold} \rangle$, is computed using a 35- and 40-snapshot data set, then the difference between these two values is taken and shown here. Purple indicates that this difference is $\geq$ 5 kcal/mol, and colors between white, gray and light purple are on the interval of [0, 5] kcal/mol. Most positions have converged (thus, mostly white), but a few outliers are found in regions that are more difficult to sample.

Figure 2.19: The average energy of each sequence, $\langle \Delta\Delta G_{fold} \rangle$, is computed over 35- and 40-snapshots. The difference between corresponding sequences from each data set is taken, and mapped out here according to position and mutation. Most positions have converged (shown in white), though some outliers can be found a more flexible regions.

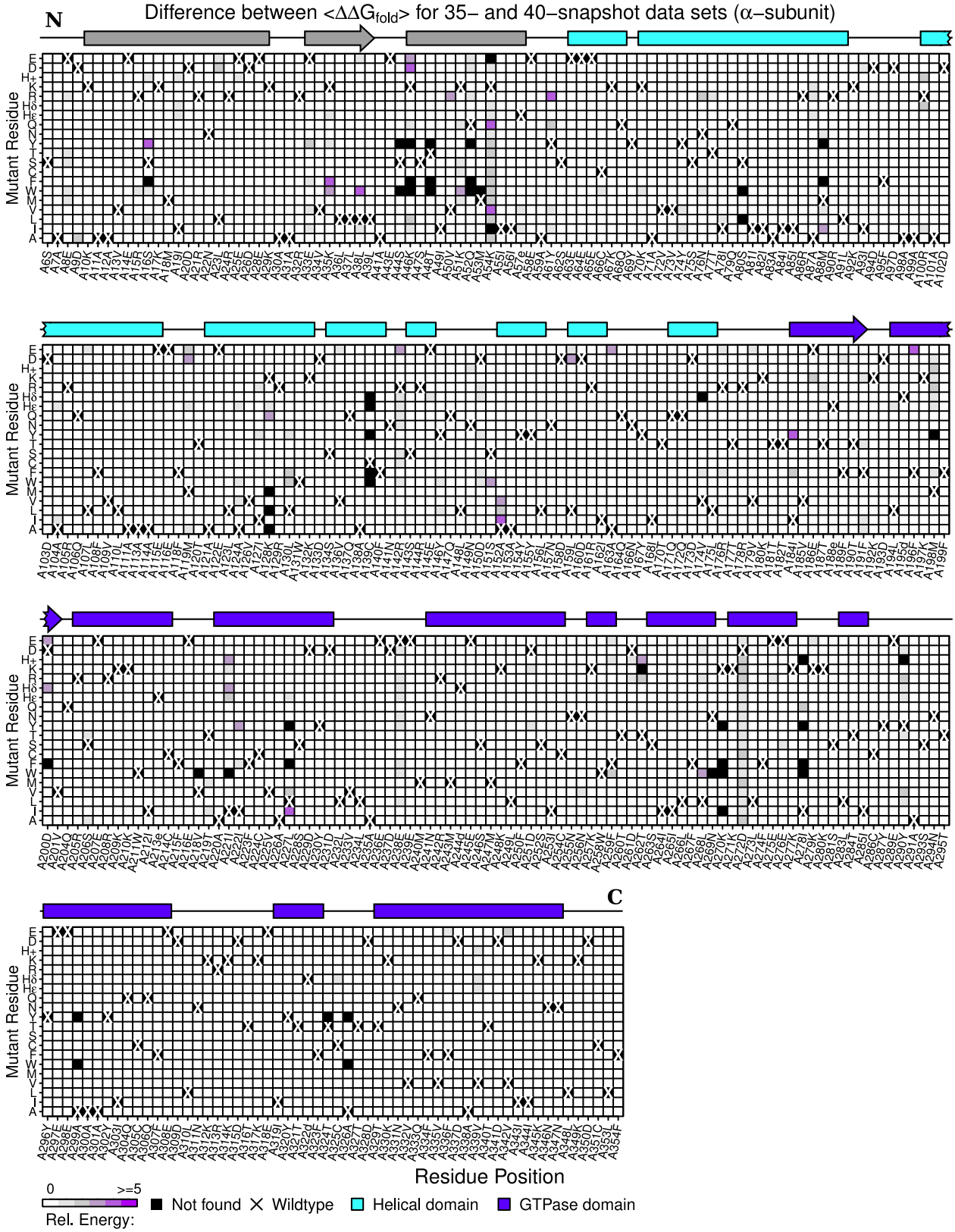

Difference between <ΔΔG_fold> for 35– and 40–snapshot data sets (α–subunit)
N
C
Mutant Residue
Residue Position
0
>=5
Rel. Energy:
Not found
Wildtype
Helical domain
GTPase domain

Figure 2.18

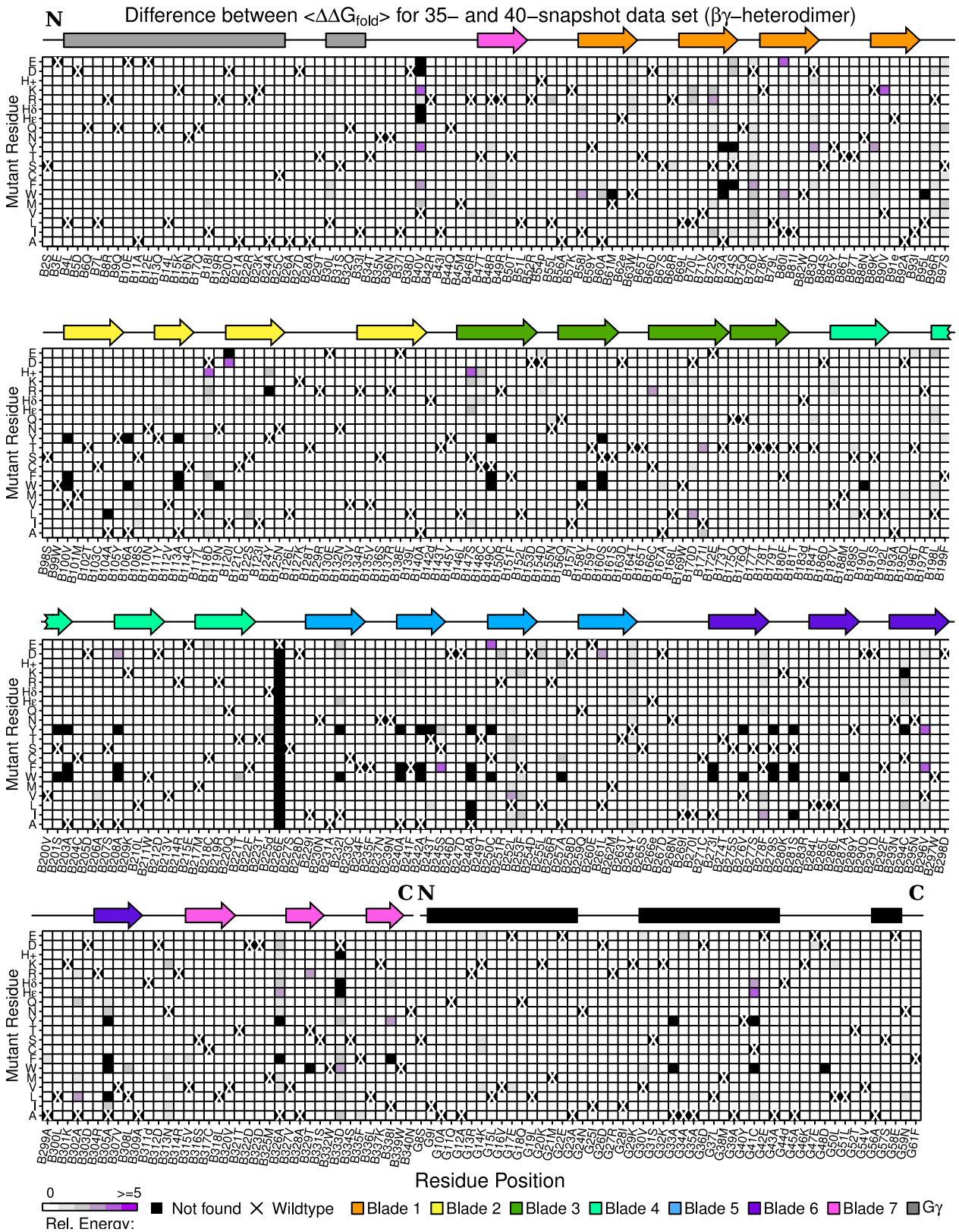

N
Difference between <ΔΔG_fold> for 35– and 40–snapshot data set (βγ–heterodimer)
Mutant Residue
C N
C
Residue Position
0
>=5
Rel. Energy:
Not found
Wildtype
Blade 1
Blade 2
Blade 3
Blade 4
Blade 5
Blade 6
Blade 7
Gγ

Figure 2.19

2.3.3 Computational alternatives to DEE/A*

Additional benefits in computational modeling of mutant structures. Alternative methods for studying wild-type contributions to protein stability, some of which require significantly fewer computational resources than DEE/A*, exist. Amino acids may be decomposed according to functional groups, for instance, so that the energy required to convert a side chain into its hydrophobic isostere can be measured, and this mutational free energy elucidates any underlying electrostatic interactions.[110, 111, 112, 113, 114] Such calculations can be completed in about 30 minutes for a single heterotrimer, while DEE/A* would require about 48 hours for the same system using the same computing cluster. The expense of using DEE/A* is well-compensated for, however—structures are simultaneously determined when finding low-energy sequences, providing explanations for how neighboring side chains can accommodate both wild-type and mutant residues by offering both visual examples of less intuitive substitutions and energetic data that can be used to organize the significance of each mutation. Energetic comparisons made with hydrophobic isosteres has its advantages in efficiency, but relies on an artificial construction that is not found in biology; DEE/A* offers practical models in its representations of actual amino acids. Furthermore, the mutant sequence space that is found by DEE/A* is very complete: nearly 99% of all possible sequences and their structures can be found for the complete heterotrimer, and from understanding how every mutant sequence behaves relative to wild type, mutational robustness can be described very thoroughly.

To illustrate the compatibility between these two kinds of calculations, and their differences, the mutational free energy of all positions involved in Gα–GDP interactions were compared to the sequences from DEE/A*. Each aspect of fitness was treated independently for assessment; the number of stable states found (defined by energetic cutoff) and distribution of mutant sequence energies were compared to mutational free energies computed using hydrophobic isosteres; positions can be separated easily according to mutational robustness in this way (**Fig. 2.20**). Lys46 has negative mutational free energy, indicating

that important interactions are made by this side chain to bind GDP, but from DEE/A*, we can understand that only wild type could ever make these contributions—no other substitutions are allowed here. Conversely, we find that electrostatic contributions of Glu43 are also important in the wild type, but have all mutations are also allowed and tend to be more favorable than wild type. Lys180 can be substituted by anything to improve stability, and also makes important native interactions, however substitutions adversely affect binding interactions with GDP—all mutations are disallowed, even though wild type has little impact on binding, again demonstrating that geometry or interactions with neighboring residues play an important secondary role.

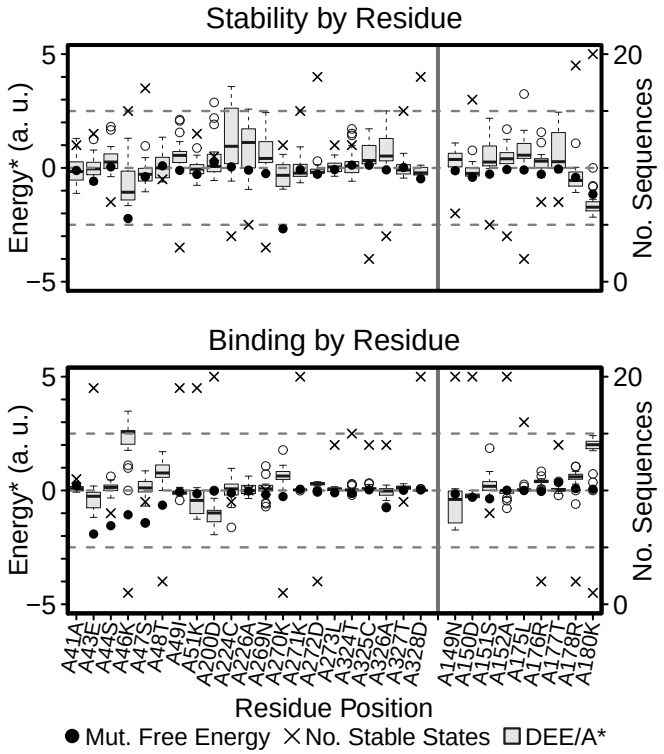


Figure 2.20: Electrostatic calculations were performed for Gα amino acids that interact with GDP. Box-and-whiskers plots represent the energetic distribution of all mutant sequences from DEE/A* at a specified position, and are overlaid onto mutational free energy data. Arbitrary units (a. u.) are (a) 6 kcal/mol for stability mutation free energy, (b) 1.6 kcal/mol for binding mutation free energy and (c) 16 kcal/mol for DEE/A* results in both contexts; the respective energetic ranges are thus (a) [-30,30] kcal/mol, (b) [-8, 8] kcal/mol and (c) [-80, 80] kcal/mol. Number of sequences found at each position from DEE/A* are at most 1.5 kcal/mol above the wild-type energy.

2.3.4 Amino-acid biophysical features captured by DEE/A*

Determining frequency of substitution, e_{ij}. Similarity matrices, such as PAM or BLOSUM, reflect the rate of substituting one amino acid for another based on protein sequences that have survived evolutionary pressures. With this in mind, it is possible to consider the protein sequence space evaluated by DEE/A* in the same way: energy is the primary selection pressure, for structural stability and binding interactions, and all sequences that "survive" this criterion ($\leq$1.5 kcal/mol for $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$) These sequences are then evaluated according to how often wild-type is substituted by a different amino acid.

Typically, PAM and BLOSUM matrices are filled with entries, S_{ij}, that indicate a score on a half-bit scale, for substituting amino acid i with j such that (**Eq. 2.3**)

$$S_{ij} = \frac{1}{2} log_2 \frac{e_{ij}}{p_i p_j} \tag{2.3}$$

where p_i and p_j are the probability of naturally finding amino acid i and amino acid j, respectively, within all wild-type sequences, and e_{ij} is the observed frequency of amino acid i replacing j. In PAM and BLOSUM, symmetry between amino-acid pairings is always assumed, but pairs (i, j) and (j, i) can be distinguished in DEE/A*, so each will be considered unique. The ratio $\frac{e_{ij}}{p_i p_j}$ is used to evaluate how closely the the observed probability, e_{ij}, is to the expected (theoretical) probability of finding i replacing j, $p_i p_j$. By rearrangement of **Eq. 2.3**, each score can be transformed into a representative e_{ij} value:

$$\begin{aligned} log_2 \frac{e_{ij}}{p_i p_j} &= 2S_{ij} \\ e_{ij} &= p_i p_j 2^{2S_{ij}} \end{aligned} \tag{2.4}$$

Wild type $G_i\alpha_1\beta_1\gamma_2$ can be used to define the probability of finding each type of amino acid naturally (the p_i and p_j terms, **Fig.2.21**). These values can then be used to convert

PAM and BLOSUM scores into the appropriate e_{ij} terms can be computed using **Eq. 2.4**. Meanwhile, e_{ij} is computed from DEE/A* data by counting the number of sequences in which amino acid i is substituted by j, after applying the energetic cutoff for evaluating fitness. These two variations of e_{ij} can then be used for comparison.

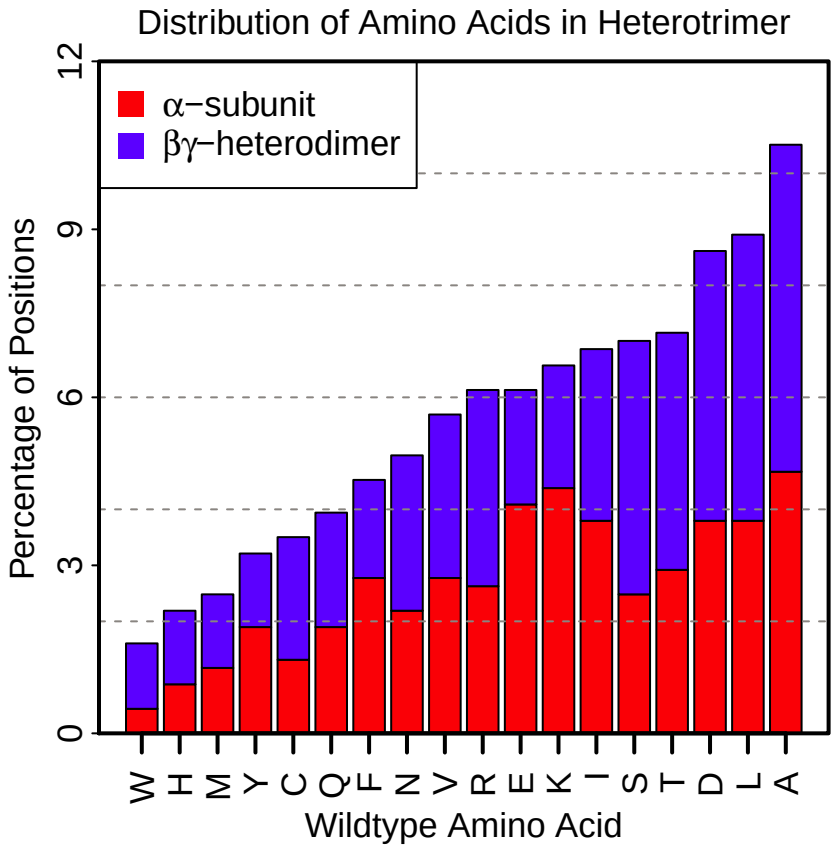


Figure 2.21: Distribution of all amino acids in the wild-type sequence is shown. The proportion found in Gα and the $\beta\gamma$-heterodimer is shown in red and blue respectively.

Correlation between DEE/A* with PAM and BLOSUM. A strong correlation exists between the expected values from DEE/A* and those derived from either PAM or BLOSUM. Protein fitness may be considered as any combination of stability and binding interactions, and the proper balance of both fitness aspects in nature is unknown. By taking different proportions of each for comparison with PAM120 and BLOSUM62, $\alpha\Delta\Delta G_{fold} + \beta\Delta\Delta G_{bind}$ (where $0 \geq \alpha \geq 1$, $0 \geq \beta \geq 1$, and $\alpha + \beta = 1$,) we best combination of stability and binding interactions can be determined (**Fig. 2.23** & **2.24**). Many

versions of PAM and BLOSUM matrices exist, and the number indicates how divergent the sequences should be for proper evaluation: high numbers in PAM indicate more divergent sequences (increase in "PAM distances"), while the number in BLOSUM reflects the approximate percentage of sequence identity between all sequences being compared. PAM120 and BLOSUM62 are common default matrices for sequence alignments when the evolutionary relationship between sequence is not obvious: PAM120 (the analogous matrix for BLOSUM80) and BLOSUM62 are appropriate when the evolutionary distance between sequence is not too disparate, at about 80 and 62% respectively.

When either PAM120 or BLOSUM62 is used, correlation with DEE/A* data increases as the proportion of $\Delta\Delta G_{fold}$ increases, until it surpasses a maximum threshold for α. Overemphasizing the importance of stability (a very large α value) will begin to decrease the correlation with PAM and BLOSUM: $\rho^2 \approx 0.8$ decreases to $\rho^2 \approx 0.7$ for BLOSUM62, while $\rho^2 \approx 0.7$ decreases to $\rho^2 \approx 0.65$ for PAM120. This suggests that structural stability is a major contribution to overall fitness that enables other functions required in protein fitness, but overemphasizing its importance (increasing α and decreasing β) will also deviate from the balance found in existing protein sequences. An even distribution between both aspects of fitness ($\alpha = 0.5$ and $\beta = 50$, **Fig. 2.22**) was used for analysis, because ρ^2 values are difficult to distinguish in any qualitative way when $\alpha \in [0.5, 0.8]$.

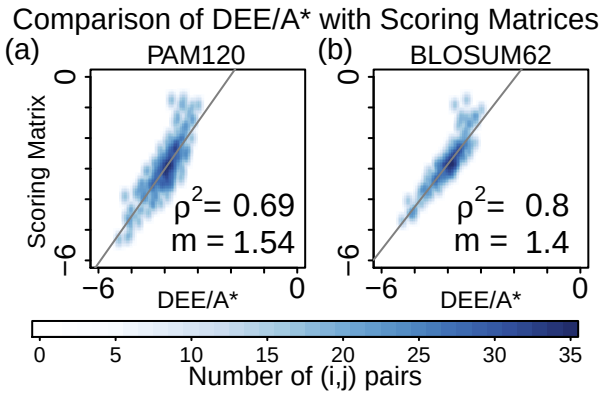


Figure 2.22: Expected frequencies of substitution for any (i, j)-amino acid pair were computed based on the number of mutant sequences satisfying the 1.5 kcal/mol cutoff. Amino acid probabilities from wild type provide a basis for deriving substitution rates for comparison to PAM120 and BLOSUM62.

Significance of DEE/A* correlation with PAM and BLOSUM. Although these correlation coefficients seem promising, the relevance of these values can be significant if this relationship is better than what can be found from randomized data. Random samples were considered by permuting the entries in PAM120, permuting the entries in BLOSUM62 or generating a substitution matrix using a uniform distribution bounded by the maximum and minimum value of both PAM120 and BLOSUM62. Robustness of the correlation between DEE/A* and randomized data was evaluated by taking 250, 500 and 1000 random samples. The worst correlation is between DEE/A* and completely random values in a fictitious substitution matrix ($\rho^2 \approx 0.2$), and better correlation is seen between DEE/A* and the permuted version of either PAM120 or BLOSUM62 ($\rho^2 \approx 0.4$ and $\rho^2 \approx 0.6$), indicating that having some discrimination between frequently made amino-acid substitutions is found in using the DEE/A* protocol (**Fig. 2.25**, **Table 2.5** & **2.6**). Not all amino acids should be substituted at the same frequency, because biophysical features can vary between them, and finding this higher correlation between permuted matrices than completely arbitrary ones indicates that DEE/A* discriminates against different amino-acid substitutions. Correlation is best when the original PAM120 and BLOSUM62 matrices are compared to DEE/A*, where $\rho^2 \approx 0.7$ and $\rho^2 \approx 0.8$, respectively. A higher correlation cannot be expected between the sequences found using DEE/A* and these two standard similarity matrices: protein fitness is very complex, involving interactions beyond those found in structural stability and binding interactions, and have not been included in the computational approach, thus leading to discrepancies between the mutant sequences that are found and the substitution rates that are based on databases of existing proteins.

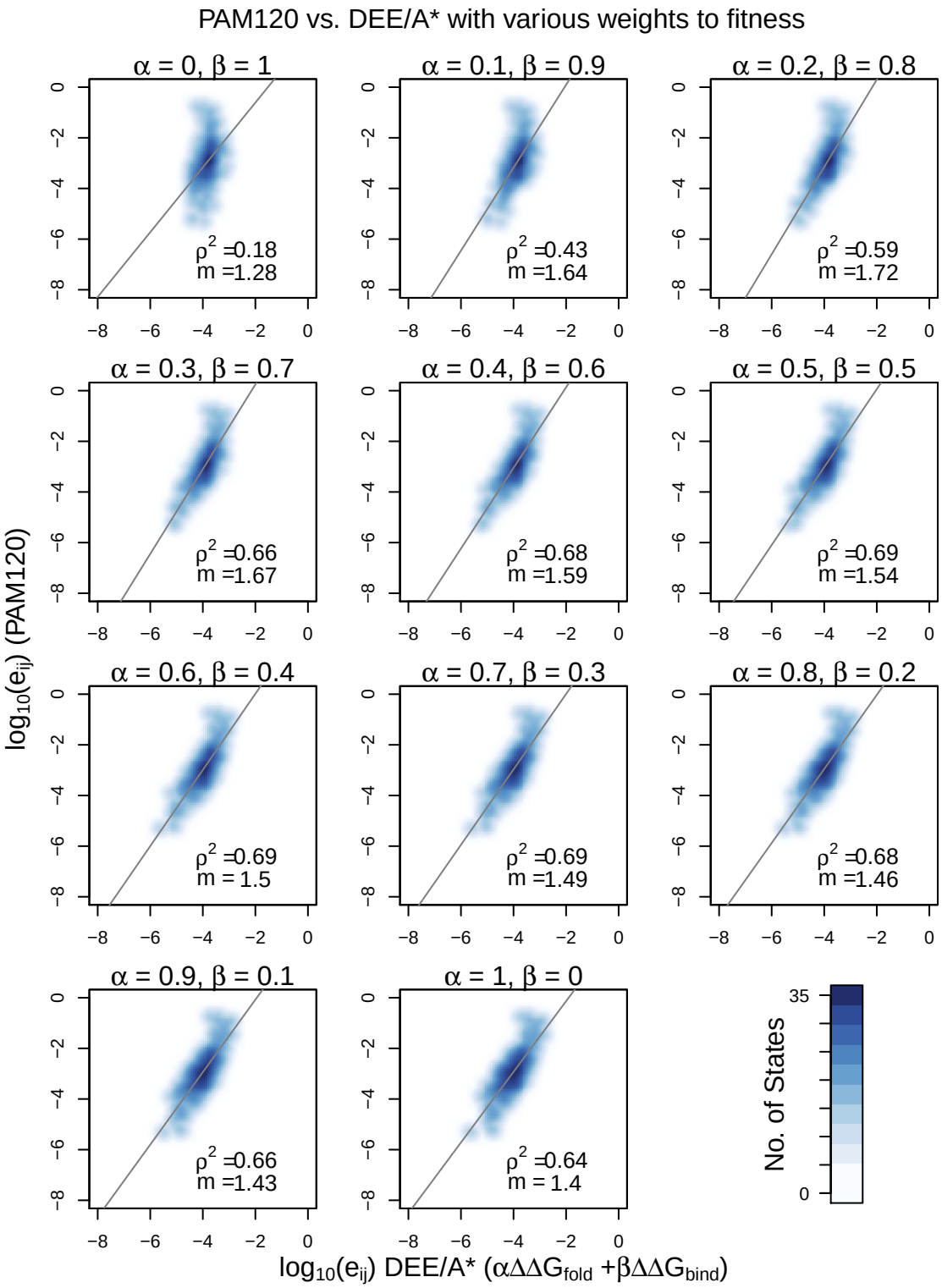


Figure 2.23: Different proportions of stability and binding were used to define the energetic criteria for survival. Correlation between the expected rate of substitution of amino acid i with j, e_{ij} , is compared between PAM120 and DEE/A* data. Pearsons correlation coefficient and slope of the least-squares fit is included. The best-fit line is shown in black.

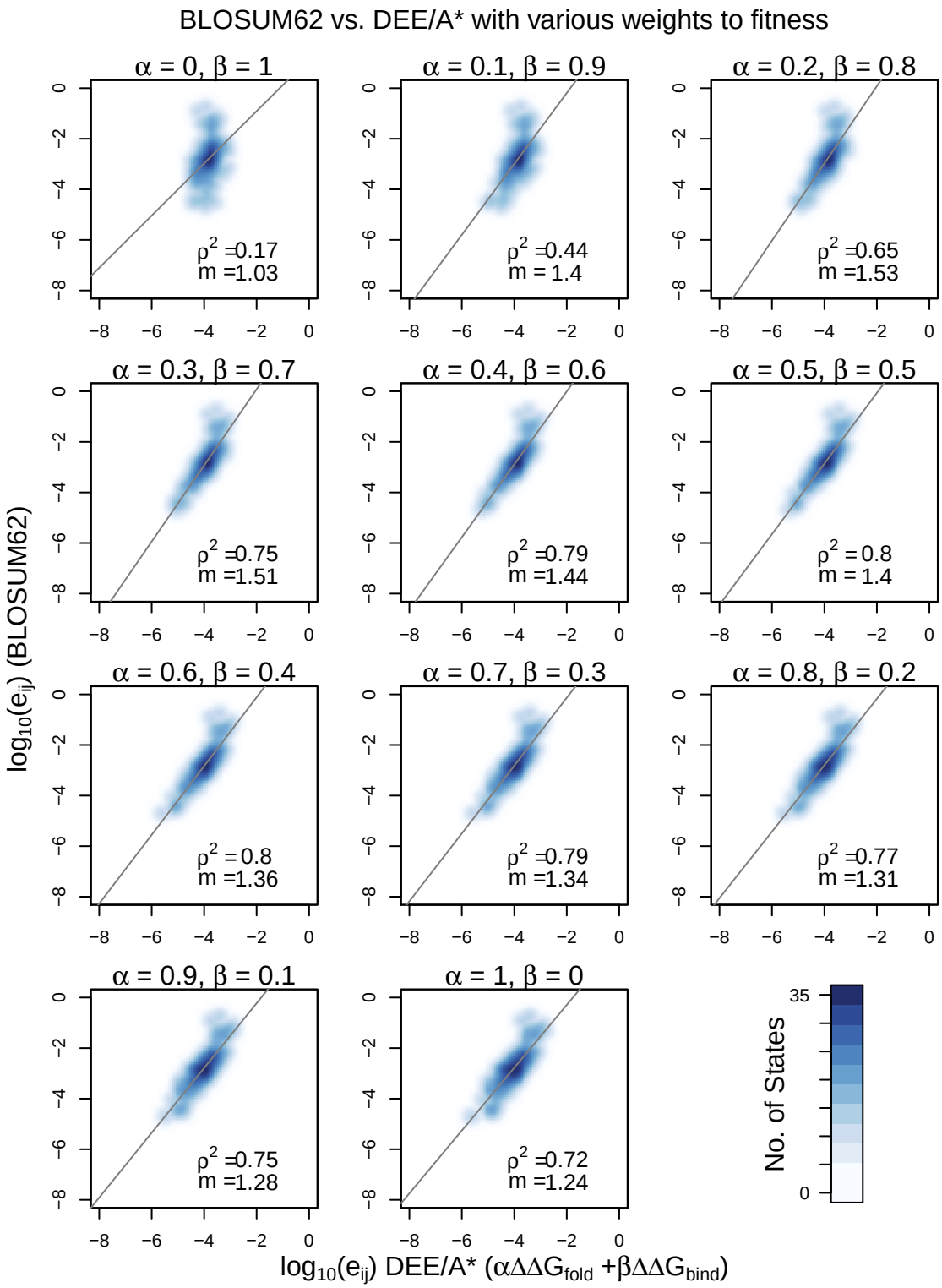


Figure 2.24: Starting with BLOSUM62 scores, the expected frequency of finding amino acid i replacing j, e_{ij} , was calculated with DEE/A* data, for different combinations of energy contribution from stability and binding interactions. Analogous values were computed from BLOSUM62, so that the two sets of substitution rates can be compared. Pearsons correlation coefficient as well as the slope of the least-squares fit is shown for each; the best fit line is also drawn.

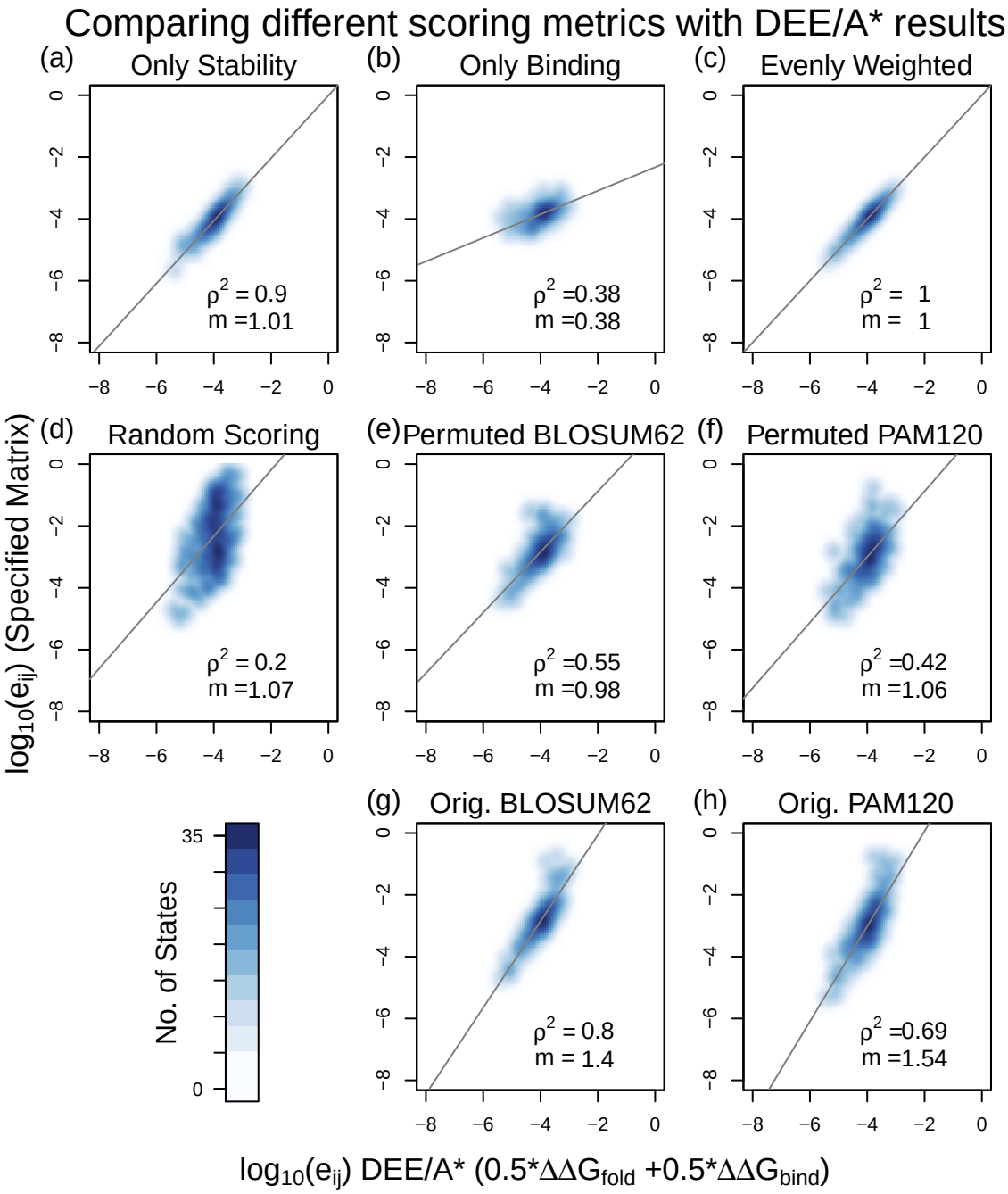


Figure 2.25: Assuming a uniform contribution from stability and binding, expected frequencies of substitution, e_{ij} , were compared to (from left to right) scores: (a) account- ing only for stability; (b) accounting only for binding; (c) in which stability and binding are evenly weighted (50–50); (d) generated from a uniform distribution, bounded by max(BLOSUM 62, PAM 120) (e) permuted BLOSUM62 matrix; (f) permuted PAM120 matrix; (g) original BLOSUM62 matrix; and (h) original PAM120 matrix.

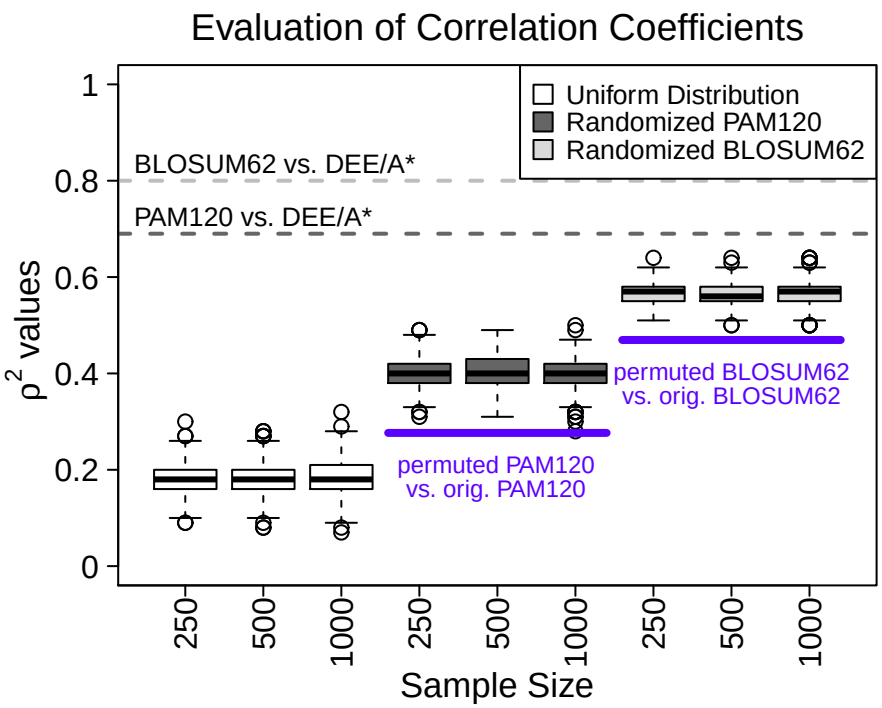


Figure 2.26: The Pearson's correlation coefficient between DEE/A* data (with 50–50 distribution between stability and binding interactions) and permuted matrix from either a random distribution (white), PAM120 (gray) or BLOSUM62 (light gray) was computed for samples of $n = 250, 500$ and 100. Distributions are generally consistent within each family of distributions. Dotted lines indicate the correlation measured between the original PAM120 (gray) or BLOSUM62 (light gray) with DEE/A* based on a 50–50 contribution from each aspect of fitness, as seen in **Fig. 2.24** and **Fig. 2.23**. Solid, indigo lines indicate the correlation between either PAM120 or BLOSUM62 with the permuted version of itself.—

Table 2.5: Average Pearson's correlation coefficient between DEE/A* and randomly generated data for various sample sizes. Number of samples given by n.

Randomized matrix	$n = 250$	$n = 500$	$n = 1000$
Uniform distribution	0.18 ± 0.03	0.18 ± 0.04	0.18 ± 0.04
PAM120 values	0.40 ± 0.03	0.40 ± 0.03	0.40 ± 0.03
BLOSUM62 values	0.57 ± 0.02	0.56 ± 0.02	0.57 ± 0.02

Table 2.6: Average Pearson's correlation coefficient between original similarity matrix and the permuted version. Number of samples given by n.

Similarity matrix	$n = 250$	$n = 500$	$n = 1000$
PAM120	0.27 ± 0.03	0.28 ± 0.03	0.28 ± 0.03
BLOSUM62	0.46 ± 0.03	0.47 ± 0.03	0.47 ± 0.03

2.4 Conclusions & Discussion

The dead-end elimination and A* search algorithms are often used to optimize a small number of sequences, but have been adapted here to explore all low-energy single-mutant sequences and their corresponding structures. In doing so, the mutational robustness of different protein domains can be quantified and compared. For the heterotrimeric G-protein, $G_i\alpha_1\beta_1\gamma_2$, the regions found to be highly sensitive to mutation in the context of stability or binding interactions are in accordance with what is well-known about this protein family through structural analysis. Unsurprisingly, mutational robustness was more prevalent in tightly packed regions, such as the core of the α or β subunits, and a greater number of allowed substitutions are found at loops and solvent-exposed areas; the allowed substitutions that retained or progressively altered protein fitness were also consistent with the behavior captured in all protein sequences found, for instance in PAM and BLOSUM, suggesting that the energetic models and algorithms used can account for important biophysical characteristics.

Additional areas for improvement include adjusting model parameters that were used for structural analysis. A longer molecular dynamics simulation would enable better sampling of protein conformations, and essentially provide better resolution of the mutagenesis data. By starting the calculations with a distance-dependent dielectric, many unnecessary calculations (i.e. due to poor steric effects or clashing) were completely avoided, and resources could be devoted to a smaller set of sequences. However, alternatives to the generalized-Born model for solvation calculations do exist. By implementing a Poisson–Boltzmann approximation instead, or perhaps an explicit solvent model, the electrostatic effects involved would be more accurate; a greater demand of computational time, however, would be the trade-off, and the efficiency of this DEE/A* protocol may be reduced. As a compromise, it would be reasonable to limit the more costly computations to smaller populations of the data set, such as favorable sequences or specific ones that can make interesting biophysical interactions in the protein.

Discrepancies do exist between what was found by DEE/A* and the frequencies of amino-acid substitution, so the correlations with PAM and BLOSUM are strong, but not perfect. Although the ability to stabily fold has a major role, overall protein fitness relies on a greater number of features beyond this and the ability to bind with known targets. Mechanisms for adaptive behavior as a response to environmental changes, for example, can also have an important role in overall fitness: allosteric changes can alter protein conformation in ways that will influence its interactions with different substrates or, relatedly, its effectiveness to modulate various effectors. A number of amino acids may be a part of this allosteric pathway, and these interactions are biologically important, but these attributes would be hard to describe as energetic functions and thus cannot be captured by DEE/A*; the environment for all single mutants in this DEE/A* protocol is well controlled, unlike naturally-occurring protein sequences that have been able to adapt against several other factors. As a part of these adaptations are also epistatic effects, the ability for alternative evolutionary pathways to arise, only in the presence of specific background mutations; if amino-acid co-variation, as pairs, triplets or larger subsets, are taken into consideration, computational modeling of mutant sequences will still fall short of perfection, but it should be possible to devise a more realistic model of amino-acid interactions and substitutions. The influence of amino-acid co-dependencies will be discussed in the following chapters.

Chapter 3

Mutational robustness of WD40-repeat proteins

3.1 Introduction

Repeating protein domains offer several advantages to protein fitness, by providing multiple regions that are highly similar in tertiary structure, despite sharing only a few signature residues in their amino-acid sequences. Gene duplication and recombination events offer an excellent mechanism for evolution, since copies of pre-existing genes can often provide a good molecular foundation for further innovation. Consequently, nearly 14% of all known proteins contain tandem repeats to some extent, and are found in organisms within all kingdoms of biology, with eukaryotes being the most common.[115, 116, 117, 118] Sequence variability between different repeating domains allow unique interactions to be made, but maintaining important, complementary ones between domains is also necessary for maintaining structural stability. Although repeating domains can be identified using approaches in comparative sequence analysis, the impact that different side chains have on structural stability and binding interactions are generally unknown, due to limitations in understanding how protein sequence conservation scores and function are related.

An important and distinctive family of repeat proteins are the β-propellers, proteins that share a special structural motif: anti-parallel β-sheets are organized in each repetition into a conformation that is similar to the propeller blade of a fan. Repeating regions are arranged in a closed, toroidal fashion, allowing the N- and C-terminal repeats to interact. Typically, β-propellers occupy a volume of space comparable to a disc or the frustum of a cone, and this geometry yields a large surface area on either side, and at the circumference, that is suitable for interacting with different kinds of binding targets.[119, 120] The innermost β strands typically form a tunnel through the protein, allowing solvent or small ligands to enter as an alternative mechanism for important binding events.[121, 122] Unsurprisingly, most amino acids in this channel tend to be excellent hydrogen-bond donors or acceptors.

The β subunit of the G-protein heterotrimer $G_i\alpha_1\beta_1\gamma_2$ is a representative β-propeller from the sub-family of WD40-repeat proteins, and the first solved structure of any β-propeller.[99] WD40-repeat proteins alone comprise nearly 1% of all known proteins across different species,[22] and engage in a variety of biological functions, including RNA processing, cytoskeleton assembly, cell-cycle regulation, vesicular trafficking, modulating metabolic activity and signal transduction across the plasma membrane.[120, 123] Overall, β-propellers are promiscuous proteins that have evolved to adopt several important roles in biology; these proteins are found in organisms that may be rather distant phylogenetically, and the primary structure of all known β-propeller have rather poor sequence identity.[124] Despite this, however, the actual fold of individual blades is often superimposable.[122]

Each propeller blade may have 4–8 anti-parallel β-sheets, and a total of 4–7 repetitions may be found, with seven arguably providing the greatest stability.[125] A blade is usually $\sim$40–60 residues in size, containing the signature Trp (W) and Asp (D) within it; a characteristic GH motif (glycine and histidine) is also known to frame the primary structure, so that a relatively consistent core length (about 23–41 residues, depending on which WD40 protein) is found between the signature GH and WD motifs of a repeat. In contrast, the sequence outside of this region remains highly variable, spanning between as few as 6 or as

many as 91 amino acids.[123] A total of four strands are found in β-propellers with seven-fold symmetry (**Fig. 3.1**). By convention, β-sheets are often named alphabetically (sometimes numerically), with strand a being closest to the center of the torus and strand d furthest out, at the edge with the most solvent exposure; loops connecting these β-sheets are called variable regions or are named according to the β-sheet that it follows it. Unlike modular proteins, in which each part of the protein may be folded from completely separate protein sequences, then assembled, β-propellers are formed from a single continuous sequence: every propeller blade has three unique β-sheets that largely defines it, but one, strand d, originates from the previous neighboring blade. Even though the four strands are within the same blade, strand d precedes all other strands in primary structure, and is connected by a hairpin to strand a, and this association between shared β-sheets between propellers contributes to the stable, circular shape of the protein.[121, 123]

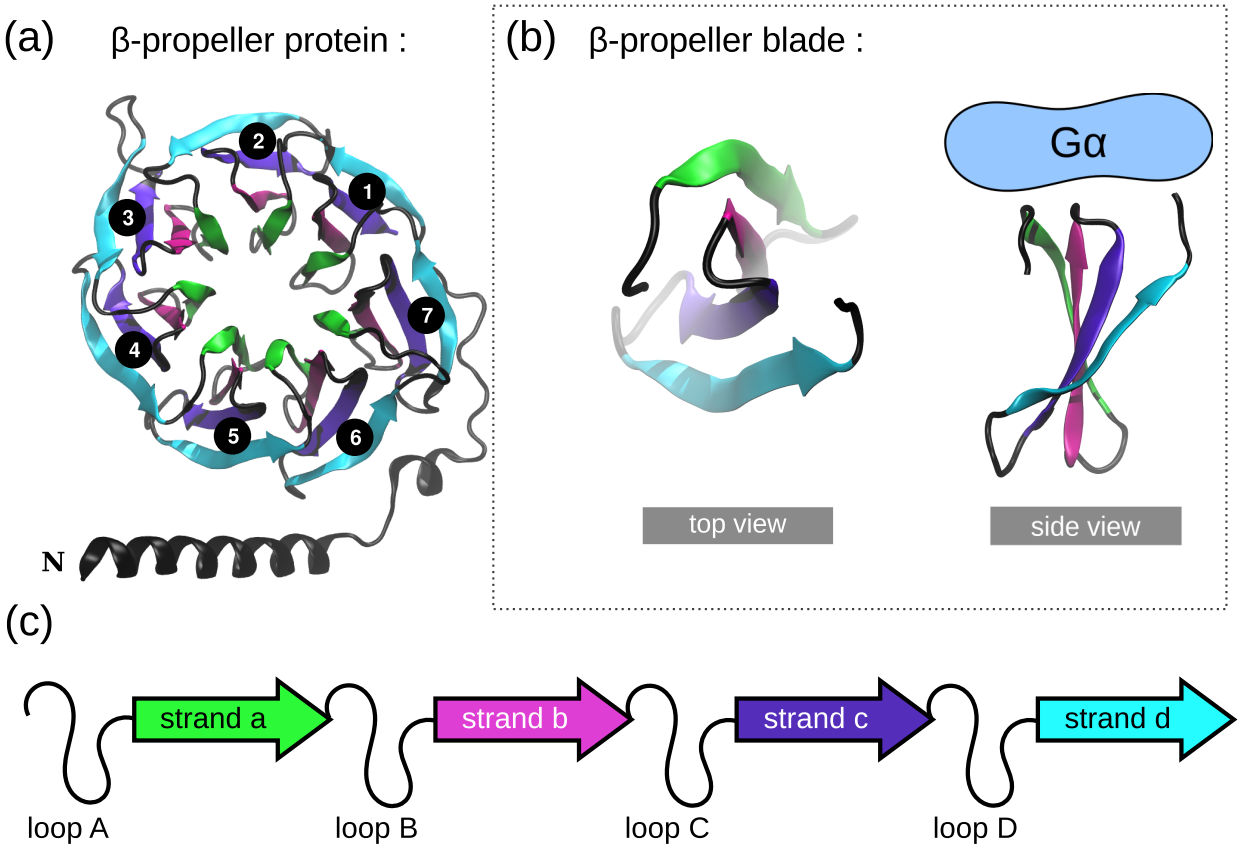


Figure 3.1: The β-subunit of a G-protein heterotrimer is a member of the WD40-repeat family, a subfamily of β-propeller proteins. In total, there are 7 repeating units, shown in (a), and each repeating region has 4 anti-parallel β-sheets, colored here according to proximity to the center of the protein. (b) These β-sheets twist relative to the tunnel of β-sheets in the middle (green) with an increasing angle to form propeller-like blades, and are arranged in a radially-symmetric fashion; this all-β fold is ideal for binding targets, such as $G\alpha_{i1}$ by $G\beta_1$. For reference, the side for $G\alpha$ interactions is shown as a cartoon. (c) By convention, β-sheets and their preceding loops are labeled alphabetically according to the location of each strand, beginning with the inner-most β-sheet, strand a, then moving outwards to strand d.

3.1.1 Known stability in β-propellers

Overall structural rigidity. The radial symmetry of β-propeller proteins provides a promising basis for studying repetitious elements that contribute to overall protein fitness, but the continuous nature of a toroidal shape cautions that small breaks in a given pattern may be necessary to maintain it. A number of stabilizing regions on β-propellers are well-known,[126] offering a few points of reference for any computational mutagenesis study. The structural dynamics in RACK1, a seven-bladed β-propeller, have been established recently through hydrogen–deuterium exchange, revealing that strands b and c are the most stable; hairpins between β-sheets may potentially be stable, but this varies across different proteins in the RACK1 family.[127] The general integrity of this protein, and its many roles in biology within many different species have suggested fold conservation, motivating several advances towards protein-structure prediction using homology modeling and related methods.[128]

The DHSW motif. A well-known mechanism for stabilizing individual propeller blades come from a near-linear arrangement of amino acids that span the pitch of some WD40-repeat proteins through a series of hydrogen-bond interactions: the Asp-His-Ser/Thr-Trp (DHSW) tetrad (**Fig. 3.2**). A given WD40-repeat protein usually has at least one DHSW tetrad, but may have up to 6 or 7.[129] Statistical analyses of WD40-repeat proteins, both from known structures and sequence alignments of potential WD40-repeats, have suggested that this motif is important to the structural integrity of β-propeller blades and have little or no involvement in binding events.[130] Mutagenesis studies in WDR5 (a 5-bladed WD40-repeat protein) have suggested that the folding energy of a tetrad may be as much as -12 kcal/mol; furthermore, there are 5 such tetrads in WDR5, one in each blade, leading to speculation that the unusually stable, cooperative and conserved interactions may potentially lead to protein unfolding in their absence.[129, 131]

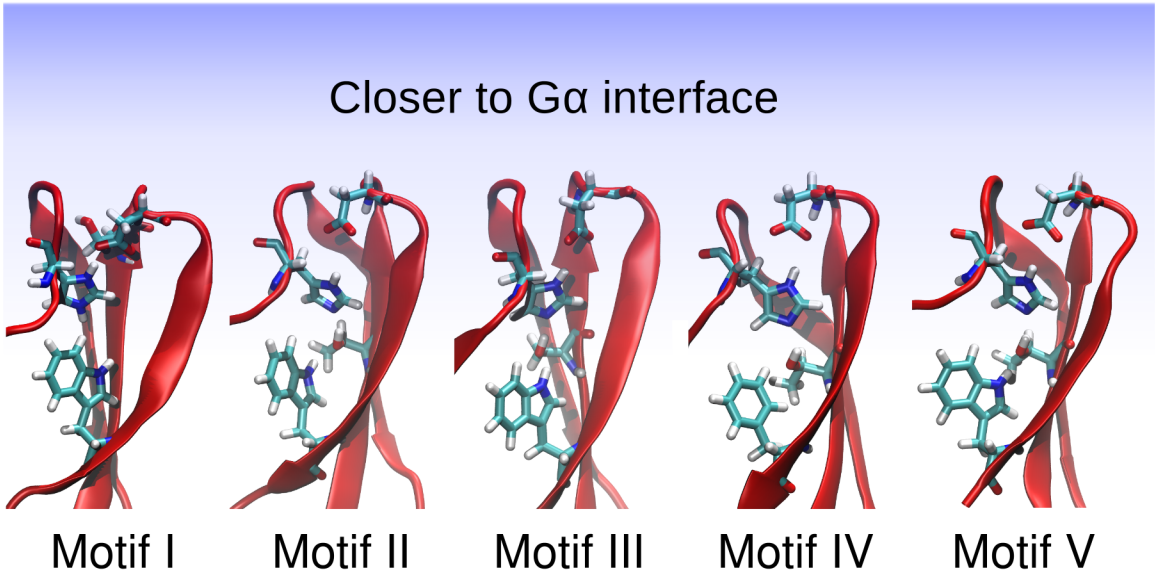


Figure 3.2: In β_1, there are 5 DHSW motifs in total: His$^+$54–Ser74–Asp76–Trp82 (motif I, found in blade 1), Hisδ142–Thr159–Asp163–Trp169 (motif II, found in blade 3), Hisδ183–Ser201–Asp205–Trp211 (motif III, found in blade 4), Hisδ225–Thr243–Asp247–Phe253 (motif IV, found in blade 5) Hisδ311–Thr329–Asp333–Trp339 (motif V, found in blade 7). DHSW motifs typically span across the different β-sheets in a given propeller, as shown here, forming a hydrogen-bond that is important to structural stability. The amino acids near the top of this page are closest to the binding interface with Gα_{i1}, rendered in blue for reference.

3.1.2 Challenges in studying repeat proteins

Sequences from repeat protein families are known to have repetitive regions in tertiary structure, even though amino acid sequences may vary greatly within primary structure; the problems that are already present in using comparative sequence analysis techniques are shared by both non-repeat and repeat protein families. In the case of the latter, however, appropriate definitions for a suitable alignment become less clear when the number of repetitions are not uniform between all proteins (**Fig. 3.3**). Similarly, it is also expected that a number of amino acids will be highly conserved, contributing to the motifs shared by all proteins within a given repeat family, and thus how signature, conserved residues should be paired together in an alignment is also unclear.

Structural repetitions, on the other hand, have several advantages in addressing the protein-folding problem. Proteins that are rather heterogeneous or globular in structure often

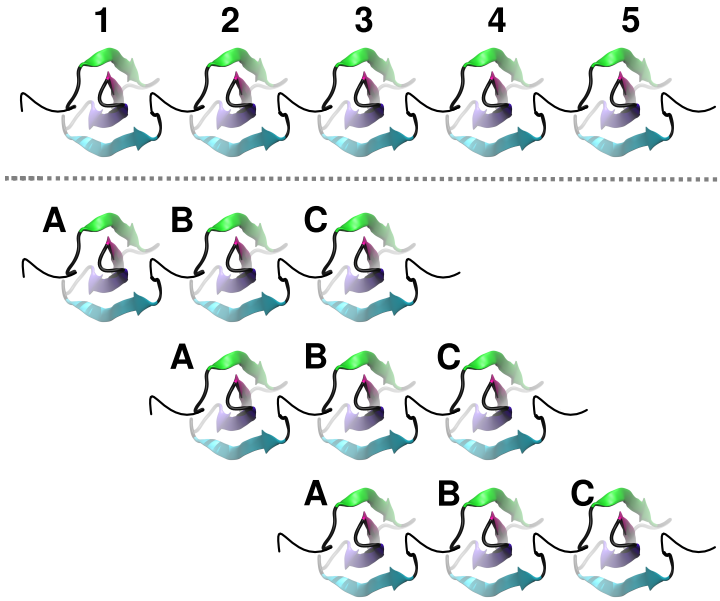


Figure 3.3: Suppose two proteins from the same repeat family need to be aligned. Appropriate definitions for comparing repeating regions becomes unclear when the number of repetitions are not the same. In comparing hypothetical proteins, one with 5 repeats (top) and the other with 3 (bottom), several possibilities exist in an alignment: (A, B, C) may align with regions (1, 2, 3), (2, 3, 4) or (3, 4, 5). By allowing gaps in the alignment, additional options would also be introduced, further complicating the problem.

rely on interactions between sequence-distant residues to stabily fold, which can often complicate analysis of protein folding kinetics and energetic variations found in experiment.[132] Cooperativity, though also found in globular proteins, is fundamental to the folding of repeat proteins. Formation of local secondary structure occurs in repetition, so regular patterns of contact formation can be more readily identified, relative to the irregular architecture of many non-repeat proteins. Still, leveraging repeat proteins for understanding the protein-folding landscape has mainly focused on the behavior of linear proteins, such as ankyrin or tetratricopeptide repeats (all-α folds),[42, 133] which are reasonably small ($\sim$20–40 residues) and straightforward to handle.[134, 135, 136]

Many examples of functional analysis in β-propellers have been introduced in the last few decades, either from extensive structural studies or by comparing relevant protein sequences,[122, 137, 138] but very few have systematically evaluated the evolvability of β-propeller structures and the attributes maintaining its overall stability. Specific amino-acid

contributions towards cooperative effects have remained unaddressed, with the exception of the aforementioned structural motifs; the interactions underlying stability in this radially symmetric protein remains unclear, and this problem is addressed rationally here, using a modified version of the DEE/A* protocol introduced in **Chapter 2**. The primary focus of this research was in determining how coupled interactions give rise to a stabily folded structure, and the repetitious nature of these interactions was emphasized to illustrate the advantages of structural symmetry. Though closely related, a more detailed discussion of amino-acid co-dependency and epistasis can be found in **Chapter 4**.

The specific goals of this research include identifying important coupled interactions based on the additivity (or lack thereof) in energetic data between corresponding mutations and how they affect structural stability. Significant interactions, defined by an energetic cutoff, are then mapped onto the structure, and important patterns of interaction can be found by leveraging the symmetric nature of the β subunit. All propeller blades share a common interaction motif that can be attributed to its overall stability, and this analysis overcomes some of the existing challenges in studying repeat proteins by accounting for structural context. Comparisons are made to the well-established DHSW structural motif, to determine how mutational robustness may vary in the β-propeller protein.

3.2 Methods

The analysis of Gβ focused primarily on structural stability, $\Delta\Delta G_{fold}$, in order to assess how coupled interactions in local secondary structure and the special arrangement of β-sheets are involved in stabilizing β-propeller proteins. Results from all single mutants in the context of binding interactions between Gα and G$\beta\gamma$ (from **Chapter 2**) were first taken into consideration for validation and augmenting the DEE/A* protocol to address this problem.

Selecting wild-type pairs for co-variation. Wild-type pairs were chosen based on a distance cutoff: at least one unique pair of atoms between the two positions must be

within 5 Å of each other, excluding the backbone atoms. Instead of using the complete 40-snapshot data set that spans 350 ns, preliminary analysis in sampling indicated that a set of five conformations, each spaced 50 ns apart would be sufficient (see **Chapter 4** for details.) Approximately 1300 unique pairs were identified in each of the five wild-type conformations, and calculations for pair-wise mutagenesis followed the same procedure as the DEE/A* protocol described for single mutants: for a given pair, solutions within 30 kcal/mol of the most energetically favorable conformation were kept for analysis.

To help validate important patterns in the double mutants, triple mutants were also explored based on the interaction motifs that were found between pairs (**Fig. 3.4**). A triplet, (A_i, A_j, A_k), was chosen for mutation if at least two unique pairs of positions within it (i.e. (A_i, A_j), (A_j, A_k) and (A_i, A_k)) have an interaction found within the interaction motif (see Results). The theoretical sequence space increases exponentially from 400 sequences to 8000, however, and thus the related conformational space that DEE/A* searches rapidly increased as well. Due to greater computational costs, a lower energetic cutoff of 15 kcal/mol was used—approximately 2-3 days are required to complete 50 unique pairs, while as many as 2-3 weeks may be necessary to complete DEE/A* calculations on the same number of unique triplets using the same resources. Often, the search space became very dense with several sequences within tenths or hundredths of a kilocalorie from each other, causing problems in memory allocation for the A* search algorithm. The lower 15-kcal/mol energetic cutoff helped maintain a tractable problem size, and most triplets completed within 3–10 days of computational time. Additionally, the search space was limited to a single wild-type backbone conformation instead of using all five. This is far from ideal for a fixed-backbone approach, but having a reduction in sampling allowed a broader sequence space to be established; while this trade-off in breadth for depth is not optimal, it provides a foundation for finding general features of amino-acid co-variation in Gβ.

Defining a reference β-propeller blade. A representative propeller blade was established according to the maximum sequence length of all individual β-sheets and loops

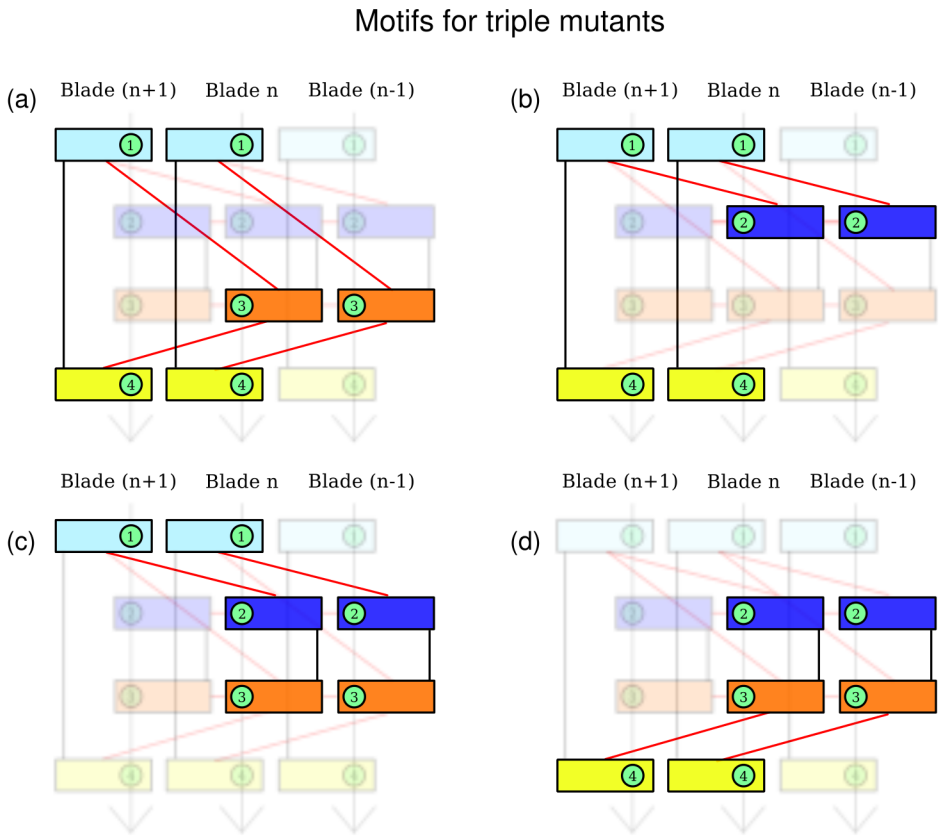


Figure 3.4: Pairs from strand a often have interactions that repeat on all seven blades and were thus chosen for mutations as triples. Triplets involving the first four residues on strand a were used, and follow the schematic representations shown here. Dashed lines in red represent inter-repeat interactions, while dashed black lines indicate intra-repeat interactions that were found in double mutants. (a) High-frequency interactions may be found between all three unique pairs within a triplet, but often only two unique interactions were pre-established, as in (b), (c) and (d).

in this particular β subunit; the greatest structural variations were at the loops and ends of β-strands, so some consistency between repeats and corresponding secondary structure is offered by aligning the sequences in this way. Secondary structure was assigned in accordance to the PDB entry for 1GP2.[99] At most 8 residues were found in these secondary structures, and this was used to build a reference motif that spans 64 positions, 8 for every strand or loop (**Fig. 3.5**). In doing so, this facilitated the comparison of corresponding residues and positions. Lengths may vary in some strands or loops, and thus sampling towards the end of a given strand or loop may be variable between all seven blades. With this approach,

some caution should be exercised: as demonstrated in **Fig. 3.5**, known sequence motifs are sometimes offset by one position from one repeat to another.

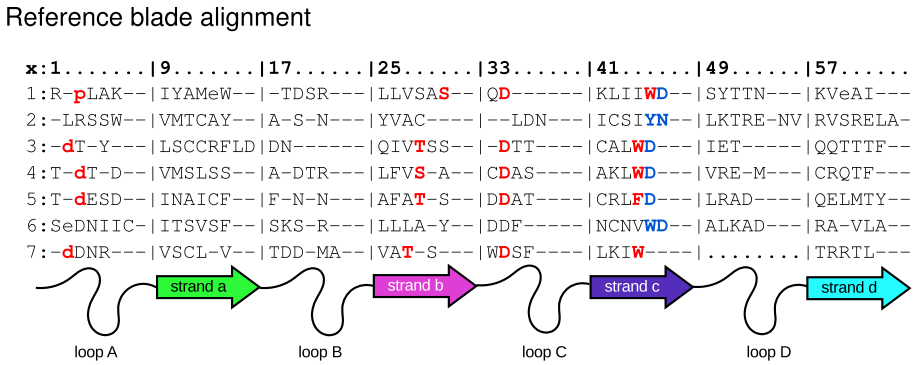


Figure 3.5: Comparison of primary structure for every propeller blade in Gβ was aligned relative to the N-terminus of each section of a repeat. Column x identifies which repeat the sequence comes from, and the numbering across the top is for reference-purposes only: sub-dividing each region by loop and strand, up to 8 positions allowed in each, yields a 64-residue sequence. Dashes indicate no mutable position was found (i.e. Pro, Gly or simply no sequence before the next loop or strand begins.) The DHSW motif is highlighted in bold red, and the signature WD motif for this family is shown in blue. The WD motif shares the same Trp as DHSW.

Identifying coupled interactions within a given pair. Every unique pair of positions, (A_i, A_j), was evaluated for additivity between corresponding energy terms. In accounting for epistasis,

$$\Delta\Delta G^{i,j}_{fold} = \Delta\Delta G^{i}_{fold} + \Delta\Delta G^{j}_{fold} + \varepsilon_{epi} + \varepsilon_{flex} \tag{3.1}$$

Here, ε_{epi} accounts for coupling (epistasis) between the amino acids at position i and j, while ε_{flex} is used to describe energetic change due to an increased region of allowed flexible side chains. Wild type is a possible choice at either position of the pair, and if chosen at one position and not another, the sequence is equivalent to a previously found single mutant; energetic differences are possible, however, because more side chains may re-orient themselves around the two chosen positions. The ε_{flex} term is essentially the average difference in energy for the same single-mutant sequence, $\langle \Delta\Delta G_{fold}(A^*_i, A_j) - \Delta\Delta G_{fold}(A^*_i) \rangle$ within the pair and single amino-acid replacement contexts. It is important to note an overestimation in ε_{flex}

may be possible, and details of these calculations are found in **Chapter 4**.

If two positions were coupled, it is expected that ε_{epi} be non-zero. Energetic cutoffs may be applied to distinguish stronger signals from weaker ones, but have not been used in this analysis to take all possibilities of coupling, such as a beneficial or disfavorable paired interaction, into consideration. To account for possible numerical errors, though, a generous cutoff of 0.5 kcal/mol was used—if the absolute value ε_{epi} was at this threshold or greater, the interaction was considered significant and the pair coupled. An interaction matrix was made for each of the five snapshots to determine the magnitude of ε_{epi}, then an unweighted average over the five matrices was computed to determine coupled interactions that were consistently found in all five conformations for each unique pair. Submatrices were then derived from this for each of the seven blades, and organized according to reference-blade position; inter-repeat interactions were then separated from intra-repeat interactions.

3.3 Results

Determining energetic cutoff for coupled interactions. A standard definition for epistasis is when a pair of interactions have a non-additive effect. However, a number of transformations have been used to assess how single and double mutants compare: (1) numerical or truncation errors could lie in computing ε_{flex}, (2) how the different single- and double-mutant energies were computed by the software, and (3) any truncation errors in subtracting these terms to compute ε_{epi}. Tolerance is usually around 10^{-6} for defining machine precision in computation, but this value is not entirely appropriate here. The energetic models used in this computational protocol are not perfect, for instance, and between taking averages and differences to evaluate epistasis between two positions, there were opportunities for noise to amplify from these operations. As an alternative, the threshold values were considered in energetic terms, and based on how the set of double mutants may proportionally change when different cutoff values are chosen.

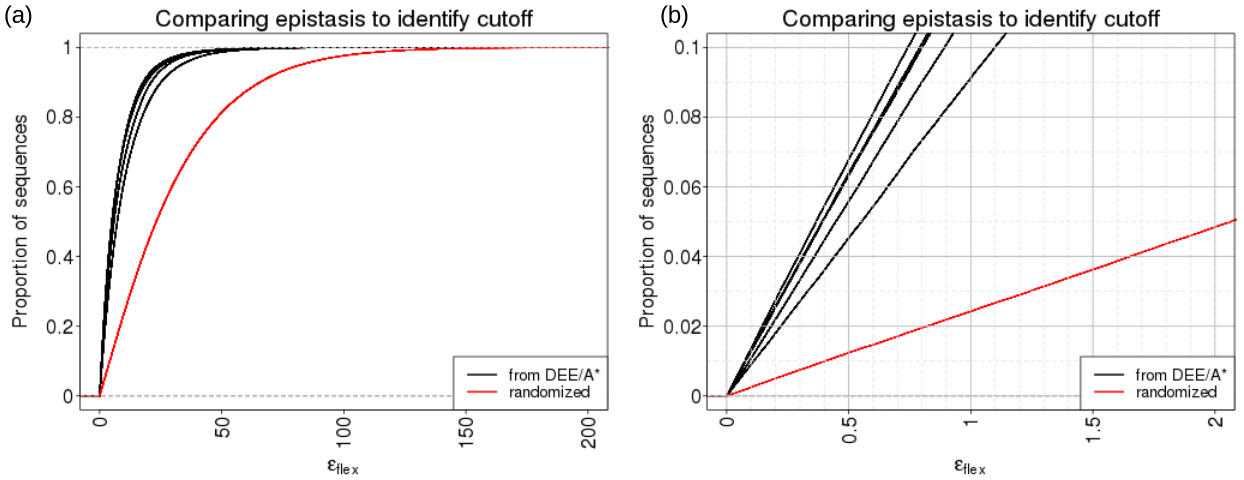


Figure 3.6: As a guideline for how different numerical cutoffs for ε_{epi} can affect the collected data, a comparison is shown here for the DEE/A* sequences (in black, about 500,000 sequences in each distribution for a given snapshot), and a randomized version of the data (500,000 sequences chosen from any of the conformations, and energy data for single and pairs were then randomized) to show the lower bound of how ε_{epi} could behave (red). The data are shown in full as cumulative distribution functions (a), and a closer look at the cutoffs is found in (b).

The influence of different cutoffs was evaluated first for each snapshot: the population of ε_{epi} terms are illustrated in **Fig. 3.6** as a cumulative distribution function to gauge how much variation ε_{epi} may undergo due to backbone conformational changes. Also, a comparison was made to how ε_{epi} would change if sequences and energies were randomly drawn from the data (shown in red.) Within each conformation, about 500,000 unique sequences are used to represent pairwise mutations. For a random sample, 500,000 sequences were drawn from all five conformations, then the ε_{epi} distribution was computed as usual, after having shuffled all single-energy data, and (separately) all the pairwise energy data; this should provide a lower bound for how ε_{epi} behaves when no actual coupling is found in the sequence (**Fig. 3.6a**, in red). A thousand replicates were drawn, and the distribution for ε_{epi} were essentially identical.

The difference between the distributions based on the original DEE/A* data and the randomized version of it is significant, indicating that ε_{epi} is non-trivial and has a specific type of relationship between single and double mutations. About 1% of the data would

be removed, if the ε_{epi} threshold were at 0.1 kcal/mol, and about 4.5–7% of the sequences would be ignored if the cutoff were 0.5 kcal/mol instead; in four of the five snapshots, more than 10% of the data would be removed if the cutoff value is at 1 kcal/mol, and this seemed too substantial. Both cutoffs, 0.1 and 0.5, were tested in constructing the reference-blade interaction matrices (**Fig. 3.9**), and there was no difference in the resulting motifs. The less conservative of the two, 0.5, was used for the ε_{epi} cutoff, to help account for overestimation in ε_{flex}; if $\varepsilon_{epi}^{i,j} \geq 0.5$, then positions i and j were co-dependent on each other. Although not guaranteed, having this 0.5-kcal/mol cutoff could also reduce the possibility that detected coupled interactions were due to numerical errors from computation.

3.3.1 Mutational robustness within individual blades

The earlier analysis that was performed on Gβ in **Chapter 2** had indicated that conservation of wild-type residues was favored at most positions. The β-sheets of each blade can be separated from the rest of the data, and arranged in sequential order to compare how one strand may be different from a corresponding one of a different blade: substitutions to polar or charged amino acids tend to be unfavorable, and this is expected because hydrophobic interactions are more common in packing together β-sheets, and few positions are highly exposed to solvent. Wild-type substitutions were generally made more easily at strand a, though energetic improvements remained modest.

Trends in mutational robustness for a standard reference blade (as defined in Methods) were established in analyzing sequences belonging to the same specific strand or loop. The structural stability and binding interactions were averaged over all sequences for each reference-blade position, $\langle \Delta\Delta G_{fold} \rangle$ and $\langle \Delta\Delta G_{bind} \rangle$ respectively, and the maximum of each was taken into consideration (**Fig. 3.7**). All mutations had either a neutral or unfavorable effect, relative to the starting wild-type amino acid: WD40 repeats are a highly evolved protein family, and so low tolerance to substitution is expected at most, if not all, side-chain substitutions. Each strand revealed some disallowance of single mutants, but this was most

prominent in strand b, and other strands that form the core of the protein (strand a and c); mutational robustness was expected to be greater, because the dynamics in this region is generally invariant, compared to strand d.[123, 127] The contributions that mutations generally have on either aspect of protein fitness can be assessed with the distribution of single mutants in each region (**Fig. 3.8**). Most single mutants had an adverse effect on overall stability, with a greater proportion of unfavorable substitutions found in β-sheets than at the loops; hairpins are naturally more structurally variable and solvent-exposed, and thus had more options for allowable substitutions. Unfavorable sequences had very similar frequencies in the β strands, but tend to be highest in the three core ones (a, b, and c). Binding interactions were less likely to be affected by substitutions, since fewer positions are involved overall; compared to other regions in the blade, loop A and C had the greatest proportion of unfavorable substitutions. They are also the two hairpins that can come in closest contact with Gα (see **Fig. 3.1**), and sensitivity to amino-acid variation here was a reflection of key interaction patterns at the binding interface must be maintained.

3.3.2 Interaction framework for β-propeller stabilization

An overview of how pairs of side chains interact was used to identify important patterns of coupled effects that contribute towards stabilizing β-propeller blades. Each unique pair was evaluated for co-dependency based on whether or not the energies of corresponding sequences were additive; in the case of coupled interactions, ε_{epi} is expected to have a non-zero value. The energy for any given sequence may be positive or negative, and this sign may change when a difference is taken between related sequences to compute epistasis. No differentiation was made between antagonistic or synergistic combinations of interactions, because identifying dependency for either reason would be sufficient in revealing the framework of necessary interactions. Instead, the frequency of finding a coupled interaction between pairs of positions on a reference blade was emphasized: if an interaction is important in maintaining the circular structure of this protein, it should be common to most, if not all of the blades.

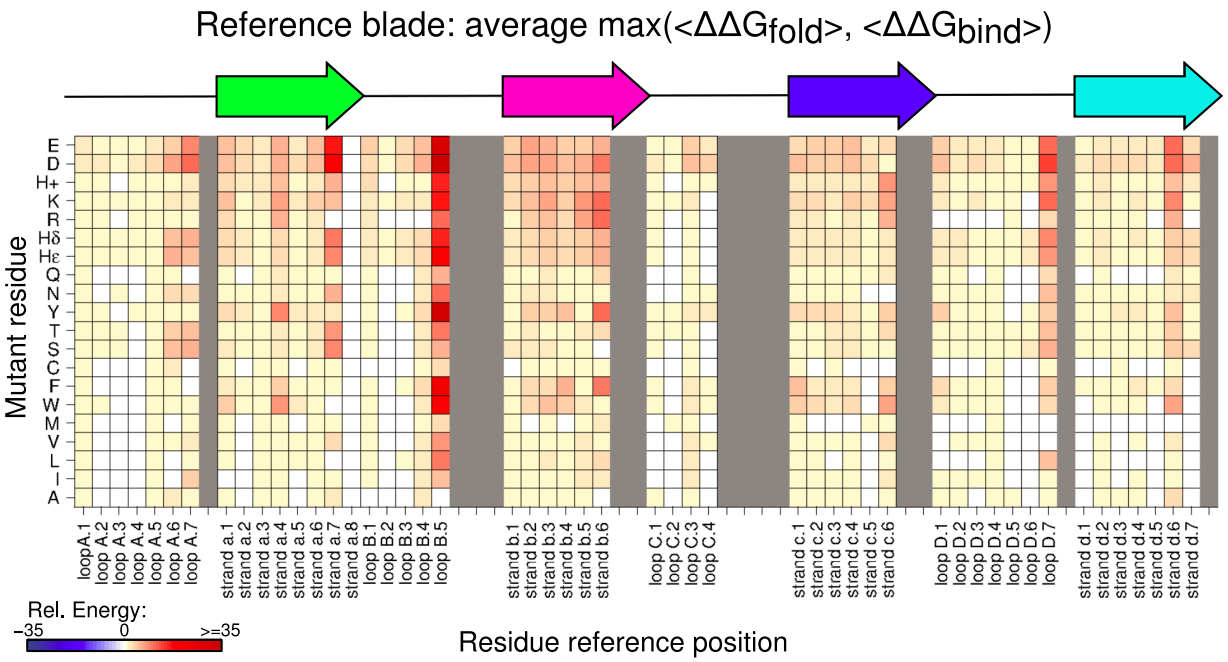


Figure 3.7: Correlation between protein stability and function is fundamental to overall fitness. The maximum of either structural stability ($\Delta\Delta G_{fold}$) or binding interactions with Gα ($\Delta\Delta G_{bind}$) is shown here for all positions in Gβ; energy for a position is first Boltzmann-weighted using the 40-snapshot data set, then an unweighted average is computed over all seven blades. Each box can represent up to seven data points, one for each blade; some boxes may be less populated if a corresponding β-strand or loop is shorter from one repeat to another. Gray boxes have no data and function here as place-holders to maintain a uniform size for all strands and loops. The energetic scale is in kcal/mol, indicated by the red–blue bar.

Intra-repeat interactions, those found between strands and loops of the same blade, were mapped in a matrix according to positions on the reference blade to determine the most frequently found coupled interactions (**Fig. 3.9**, top). Signals in this interaction matrix revealed branching off the diagonal of the matrix, due to the intercalation of amino acids on anti-parallel β-sheets, and this is a feature of any contact map that involves this secondary structure. The proximity of side chains between any pair of β-sheets was primarily responsible for the greater number of interactions between any two strands, relative to how a loop may interact with a strand or another loop. A comparison was made between how these interactions varied when all five conformations were considered as an ensemble (**Fig. 3.9a** & **b**), relative to how one conformation behaved (**Fig. 3.9c** & **d**), as a guideline for how

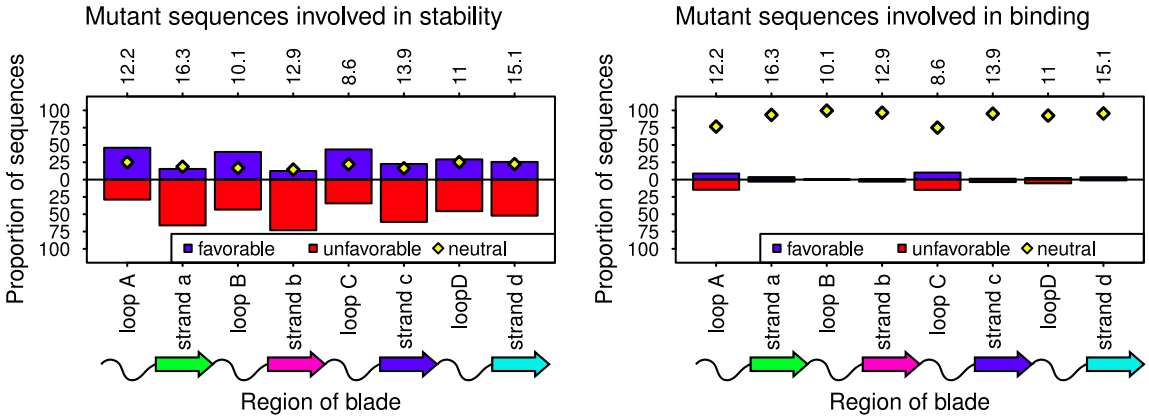


Figure 3.8: Trends in mutant sequences for each strand and loop is assessed by counting all sequences found within a given energetic cutoff. Colors represent the proportion of states that are favorable (blue), neutral (yellow) or unfavorable (red) relative to wild type, based on $\langle \Delta\Delta G_{fold} \rangle$. Structural stability is evaluated in the left panel, and binding interactions with Gα are shown on the right. As an indication of sample size, the proportion of single mutants for each strand and loop is shown on the top axis as a percentage.

backbone fluctuations affected co-dependencies. Most interactions were conserved between both sets of data, and a few high-frequency intra-repeat interactions were lost (**Fig. 3.9a** & **c**, upper triangles of each matrix); from this, structural variation was indicated to be less frequent at strand a, and more fluctuation between pairs could be found at the other strands. Most paired interactions occurred in only three or four blades, but a small number of interactions were found in six or seven of the repeats within strands a and b. These high-frequency interactions follow a well-established pattern of β-sheet side-chain interactions: positions at i and $i + 2$ are more likely to interact with each other than with positions that immediate follow in sequence, because side-chains generally alternate direction in β-sheets.[125]

Fewer coupled interactions were found between any two propeller blades (**Fig. 3.9**). Natural twisting of β-sheets in a propeller means greater solvent exposure and a larger distance between corresponding β-sheets of adjacent repeats, in contrast to the proximity of neighboring β-sheets in the strands at the central tunnel. Thus, a greater number of coupled interactions were found at strand a than at any other strand or loop, due to spatial

orientation of secondary structure. As with intra-repeat interactions, the majority of coupled interactions were found in fewer than half of the repeats; the strongest patterns of interaction were at strand a and b, often following the i and $i+2$ patterns as well.

High-frequency interactions, occurring in five or more blades, were evaluated more closely to understand the biophysical interactions within the β subunit (**Fig. 3.10**). From this graphical representation, it becomes obvious that the most highly connected positions (nodes with the most edges attached) tended to be at strand a for inter-repeat interactions. Again, this is influenced by the proximity of side chains within the central tunnel, and also implies that pair-wise coupling is significant to stabilizing the β-propeller. A number of intra-repeat interactions were also found, mostly between strand a and b, helping to anchor the gradually more solvent-exposed β-sheets to a stable framework of interactions at the center. Inter-repeat interactions were seen in either strand c or d—β-sheets at the edges of a propeller are expected to be more distant and interact less, due to a larger angle of twisting. Structural examples indicated that a series of rings can be formed from side-chain interactions: in the cases where interactions were repeated seven times, the side chains tended to point in the same direction, extending towards strand a in the next β-propeller blade; when interactions were found in six of the repeats, a second set of interactions could help compensate for any deficiencies by having side chains oriented in the opposite direction to that of the first ring of interactions.

Primary structure for the highly-repetitive coupled interactions was also taken into consideration (**Fig. 3.12**). For a given pair, many amino acids at one position were very similar, for instance, strand a at position one is either Ile, Val or Leu, and in strand b at position 4, amino acids are Ser, Cys, Thr or Ala. It was uncommon to find symmetric substitutions within a given pair: between strand a at position 6 and strand b at position 2 (A6–B2), (Leu, Phe), (Phe, Leu) and (Trp, Leu) could be found at this motif, suggesting that analogous structural environments were sometimes found between n and $(n-1)$. Interactions at the central tunnel were of particular interest (**Fig. 3.12a**) since the highest repetition

of coupling tended to be found here. In strand a, the first four positions were involved in five unique interactions, and the orientation of the side chains followed a staggered, off-set packing principle to allow this (**Fig. 3.11**): instead of a pattern between i and $i+2$, the coupled relationship was between either i and $i+3$ or i and $i+1$.[125]

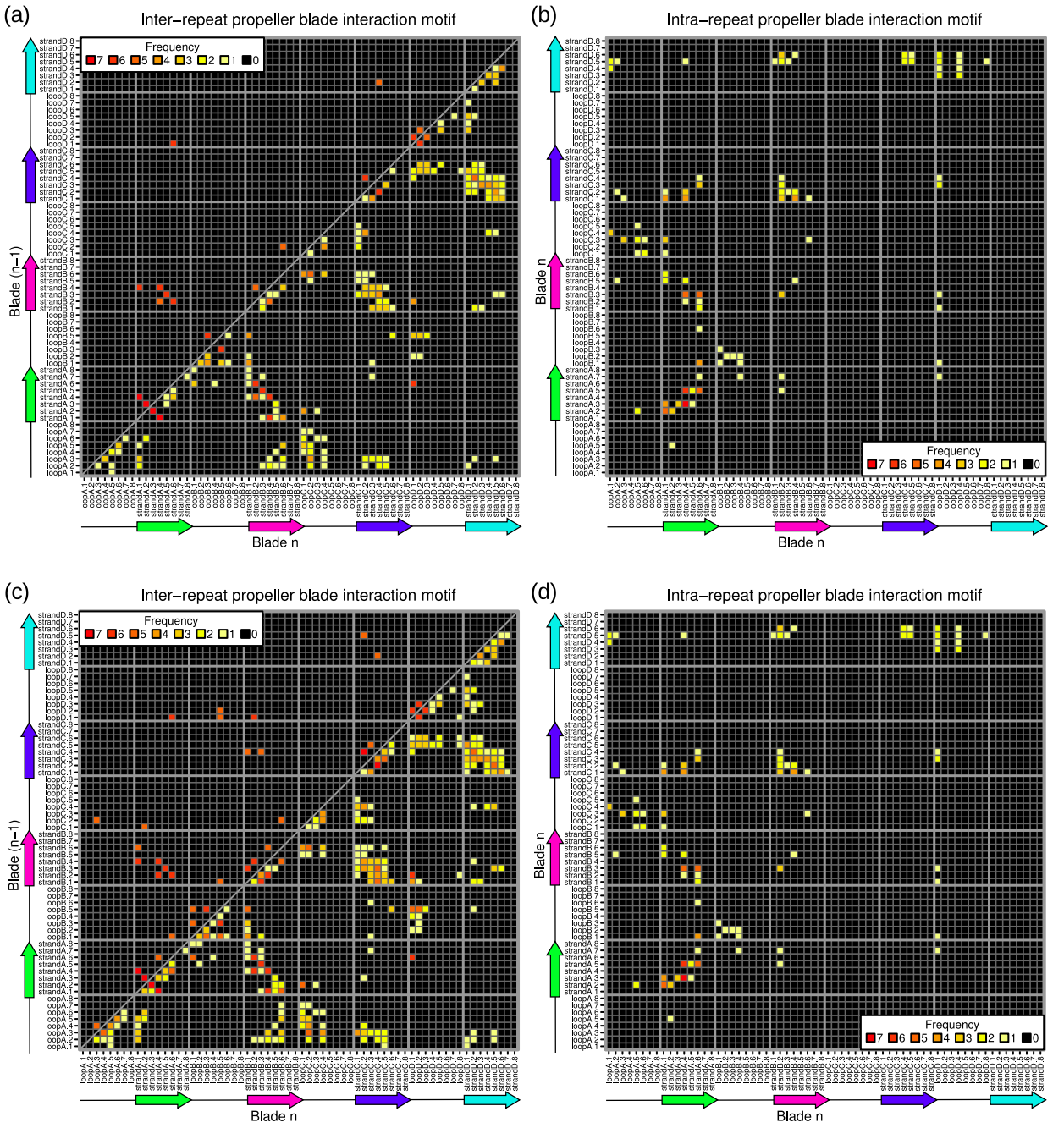


Figure 3.9: Interactions in which $\varepsilon_{epi} \geq 0.5$ kcal/mol are shown here for the reference blade. Interaction matrices are organized according to secondary structure of blades n and $(n-1)$ for intra-repeat interactions ((a) & (c)) and for inter-repeat interactions ((b) & (d)). The colors correspond to how often a coupled interaction was found out of all seven propeller blades. Interactions from the complete five-structure ensemble is represented in (a) & (b), while a single conformation was used in the maps (c) and (d). A noticeable number of pairwise interactions are missing between the two, but a number of highly-repetitive motif interactions were still conserved. (See also **Fig. 3.10**.)

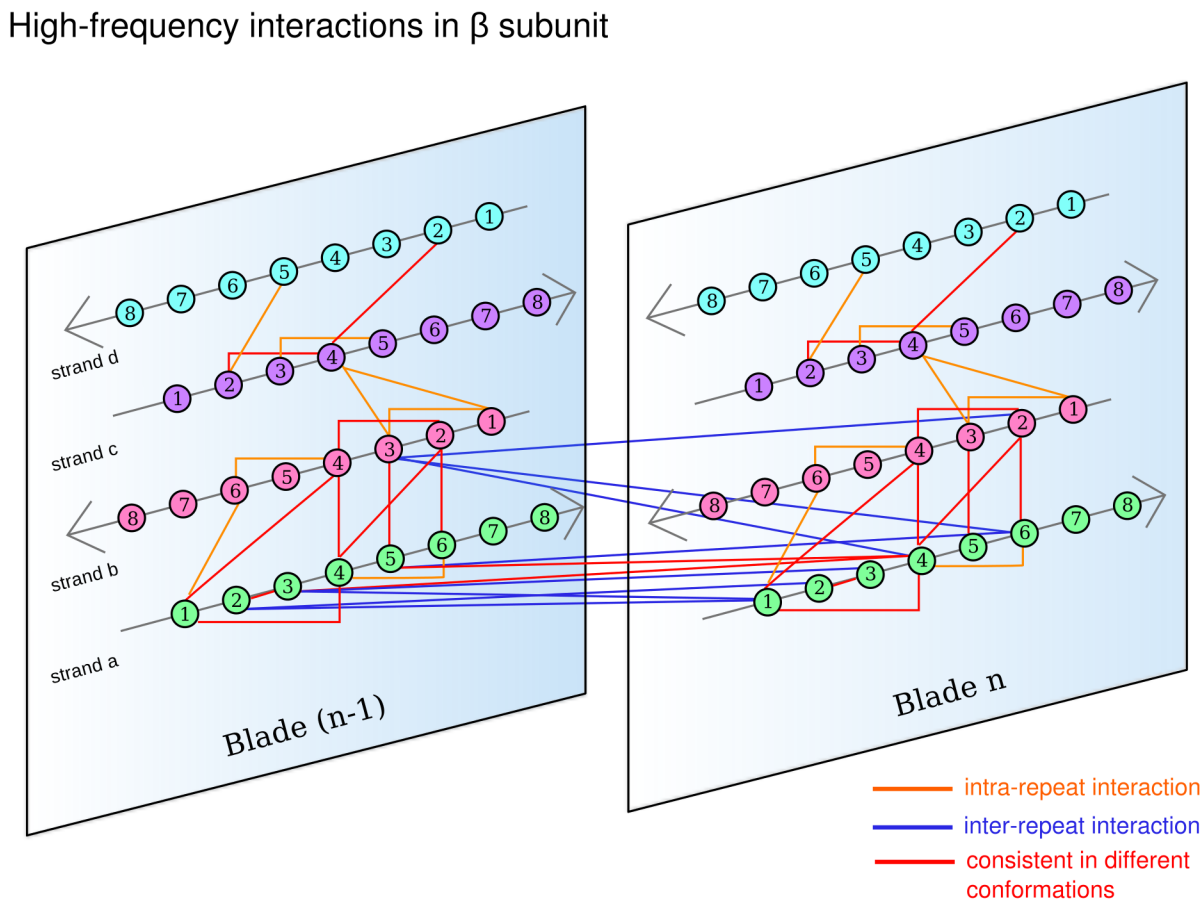


Figure 3.10: A graphical representation using nodes and edges to represent different positions in a reference blade is shown here from the perspectives of blades n and $n-1$, as an alternative to **Fig. 3.9**. An edge was drawn between a pair of nodes if a coupled interaction was found in five or more of the seven blades; inter- and intra-repeat interactions were distinguished. Interactions found in the complete ensemble (five distinct conformations) was differentiated from those found in one conformation to demonstrate how consistent some of these interactions were.

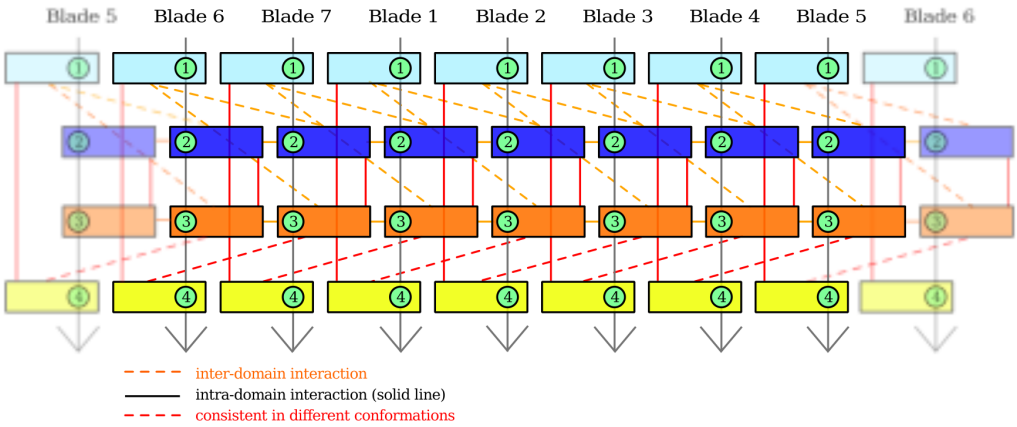
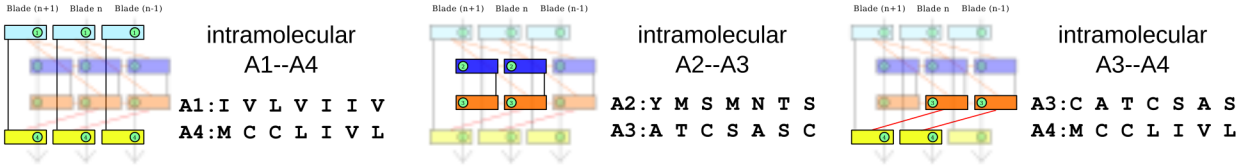


Figure 3.11: Coupled interactions within the central tunnel formed by strand a from all blades may occur as often as 6 or 7 times. These interactions were primarily on the first four residues of each strand a β-sheet, and were expected to wrap around (shown faded and blurred) to maintain radially symmetry.

(a) found in 7 repeats



(b) found in 6 repeats

intermolecular A4--A5	intramolecular A1--B4	intramolecular A4--B4	intramolecular A5--B3	intramolecular A6--B2	intramolecular C2--C4
A4:C C L I V L	A1:I V L V I I	A4:M C C L I V	A5:e A R S C S	A6:W Y F L F F	C2:L C A K F C
A5:e A R S C S	B4:S C T S T A	B4:S C T S T A	B3:V A V V A L	B2:L V I F F L	C4:I I W W T V

Figure 3.12: Some interactions were found in nearly all propeller blades in Gβ. The pairs of residues for each type of motif is shown here, and sequences are labeled according to β-strand and residue position counting from the N-terminus.

Motif interactions of high repetition tend to be robust. Signals in the reference motif were strongly correlated with the contact map of the protein, and this was expected since most positions (∼90%) have ε_{epi} values that exceeded the chosen threshold. Of course, this covers a broad distribution of coupling, from moderately weak ones to those that are very strong. As a way to differentiate how the highly-repetitive motifs (occurring in 6 or 7 repeats) behaved relative to all other unique pairs, the subset of protein sequence space for each pair was compared (**Fig. 3.14**). For each unique pair, the proportion of favorable, neutral and unfavorable sequences found was computed, and an unweighted average over all pairs was computed: approximately 80% of all double mutant sequences were unfavorable, while about 18% of them were neutral and the rest were favorable.

A comparison was made with these values for the unique pairs found in highly-repetitive motifs: 21 pairs were found in 3 motifs (found once in each blade), and 18 pairs were found in 6 motifs (each found once in all blades, except blade 7.) Nearly all of these pairs behaved differently from the average proportions found in the data set. Most of these positions had 90% of the double mutants, if not more, in an unfavorable energetic state relative to the wild type; almost none of these sequences were found to be more favorable than wild type. While most double mutants tended to be unfavorable, at least one sequence could be found to be energetically preferred over wild type in about 80% of all unique pairs within the entire data set (**Fig. 3.13**). Within the group of highly-repetitive motif interactions, the probability of finding at least one favorable state in a pair was substantially lower, suggesting that very specific native interactions need to be maintained at these positions.

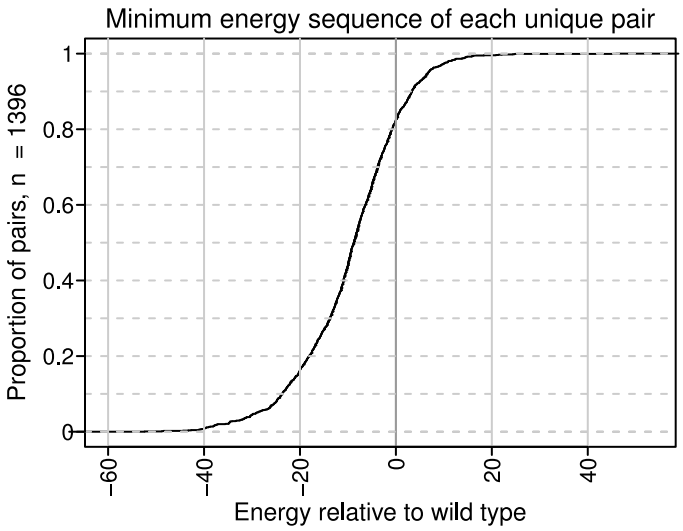


Figure 3.13: Many double mutants were unfavorable relative to wild type, but many of the unique pairs may have a few sequences that were favorable. To assess how common this may be, the minimum energy sequence of all double mutants found at a given pair was determined. The distribution of these lowest-energy sequences (relative to wild type) is shown here. Nearly 80% of all unique pairs had at least one state that was more favorable than wild type.

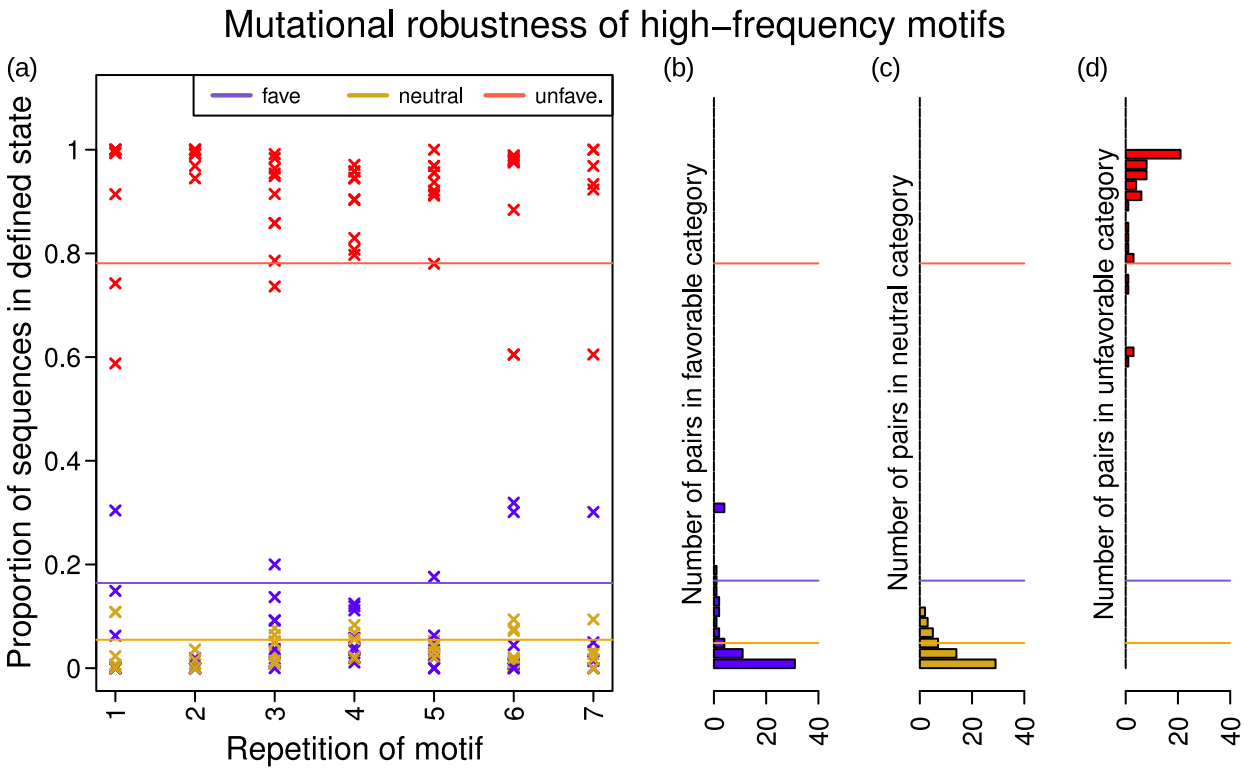


Figure 3.14: The proportion of favorable, neutral and unfavorable sequences found at each unique pair can be used for comparative analysis. The average proportion of sequences in each of these three states is shown in solid lines. In panel (a), the frequency of finding sequences in each state is shown for unique pairs found in highly-repetitive motif interactions (within 6 or 7 blades.) The population sizes for each are found in panels (b), (c) and (d).

3.4 Conclusions & Discussion

The β-propeller blade provides an excellent platform for a number of different binding targets due to its geometry. The analysis carried out here demonstrates how DEE/A* can be adapted to search over double-mutant sequence space to understand amino-acid co-dependencies in Gβ. Similar to the results found in **Chapter 2**, many substitutions to wild type were less favorable, which continued to suggest that the β-propeller protein is immutable, given its evolutionary history; while many binding partners exist for β-propeller proteins, sequence variation accounting for these types of interaction were not considered here—structural stability was the primary focus, and how promiscuity in this protein family is achieved remains unaddressed. By decomposing the β subunit into each of its propeller blades for comparison, several repetitive interactions could be found within a given blade and also between them. Only a small handful were consistently found in six or seven of the repeats, and they were generally on the strands expected to be more stabilizing. From assessing the double mutants at these highly-repetitive motif interactions, the proportion of sequences found to be unfavorable tended to be much larger than the majority of pairwise interactions studied. The repetitious nature of these interactions, and the apparent intolerance to amino-acid substitutions indicated that they share an important role in β-propeller stabilization.

Although mutational robustness was prevalent in several pairs of interactions, the basic framework for structural stability in Gβ may be more extensive. A more thorough analysis on protein sequence space could be used to assess this, because few limitations were set in biasing against different types of epistatic effects. Mutations as single substitutions could be antagonistic as a pair, for example, dampening mutational effects, or they may be synergistic, amplifying the change in fitness relative to wild type. This distinction was overlooked, because the proportion of unfavorable sequences was already high (more than 60% in single mutants, and about 80% in pairwise mutations.) Numerically, it would still be possible to distinguish between antagonistic interactions from synergistic ones, but the qualitative effect would often be the same: one may be worse than the other, but essen-

tially they were both worse than wild type. In addressing overall structural stability, this characterization of pairwise interactions is not entirely helpful.

A second consideration to help clarify coupled interactions would be to assess the number of coupled interactions at a given pair more carefully. The analysis presented computes $\langle \varepsilon_{epi} \rangle$ for a given pair, but a better description of this average could also be used. For example, suppose pair A has positions that only a few coupled sequences with a very high average epistasis term, while another, pair B, has nearly all double mutants found, and a similar average epistasis value. In pair A, coupling could be perceived as being weaker, because fewer pairs are found to be co-dependent than in pair B, and this can be an important distinction to have in some situations. However, given the high proportion of unfavorable sequences for Gβ, this type of assessment would often involve coupling between unfavorable substitutions; it is then possible that fewer coupled pairs were found at pair A for a few reasons, including the fact that some unfavorable sequences exceeded the energetic cutoff used in DEE/A*, and they may also be coupled. For this reason, mutational robustness was evaluated here as an average over pairs that were found, and special weight was not given towards pairs with a larger coupled sequence space.

Chapter 4

Influence of coupled interactions on structural stability

4.1 Introduction

Maintaining overall protein fitness requires having concerted interactions between different amino-acid side chains within a given protein, and thus determining amino-acid co-dependencies is essential to understanding how mutational robustness can be achieved. Introducing an amino-acid substitution into a sequence to alter the current level of fitness may seem deceptively easy, but because of highly coupled interactions between different positions, a small molecular change may be highly destabilizing to the native interactions.[90, 92, 91] Protein co-variation may be considered at the molecular-interaction level, in which amino acids and their specific interactions are examined, or at the full protein-sequence level, where more attention is paid to the phylogeny of protein families and how one progressed and evolved into the next. In combination, these two aspects of evolution provide different mechanisms for understanding how sequence conservation may be related to fitness:[44] by keeping track of progressive changes in amino-acid variation, resistance to changing wild type at a given position in coordination with (or as a consequence of) some modification elsewhere can

be determined, and this provides insight on how sequences may become adaptive in dampening unfavorable changes to fitness. On the other hand, following the ancestral history of a given protein, to see how one sequence was able to evolve into another or (equivalently) which ancestors may be shared, can be potentially advantageous in distinguishing key interactions in structural stability or correct function for the family at large from interactions that are required for specificity within a protein sub-family.

Within a protein sequence space, amino-acid substitutions will yield one of three effects: either the sequence will have enhanced fitness (advantageous mutation), reduced or loss of fitness (deleterious mutation) or no observable effect is found in the phenotype (neutral mutation.) Mutations at the genetic level could alter the balance found in overall fitness, yielding destabilizing effects, but if compensatory interactions were introduced, also by mutation, it may be possible to dampen deleterious effects, restore the previous level of fitness or even beneficially enhance fitness beyond native wild-type attributes. By implementing the same techniques as those used to predict co-varying pairs, the interactions underlying these epistatic effects, whether positive (advantageous) or negative (deleterious), can also be detected.

A number of quantitative methods currently exist in disentangling the relationships between different amino acids within the same protein structure. All of them, however, depend heavily on multiple sequence alignments. The argument for developing these techniques is that sequencing happens at a much faster rate than protein structure prediction, and having algorithms to detect coupling between pairs of positions will elucidate how protein contacts need to be made, potentially helping in structure prediction. Multiple sequence alignments for assigning conservation scores is different from the alignments needed to understand protein co-evolution. Namely, phylogeny has a very important role in how sequences may be closely related, and needs to be taken into consideration in analysis to understand possible biases in sampling. When a coupled interaction is found between positions and paired in a structurally or functionally useful way, it is not always clear if the relationship

was inherited from a common ancestor, and the answer becomes less obvious when the sequences used in an alignment are highly similar. Cutoffs in sequence similarity—usually about 80-90% for an upper bound, and about 20-30% for a lower bound—becomes important for the alignment, but these values can be varied, depending on how gaps insertions may have been used. At the same time, though, these cutoffs can be seen as being almost arbitrary to biology, based mostly on statistical reasoning.

After a multiple sequence alignment is performed and an algorithm is applied to it, there are still questions of how to distinguish high-scoring pairs from lower-scoring ones. Often, numerical cutoffs are used, evaluating the top m pairs or those that satisfy a statistically-based cutoff, such as p-value, and ignoring the rest. In many ways, this is analogous to the problem of understanding conservation scores: what do the scores in between 0 and 100% actually mean? For this reason, DEE/A* provides a major advantage in being able to isolate specific aspects of fitness for assessment, by utilizing structural and energetic data, readily available from the algorithm, to evaluate how different pairs interact.

4.1.1 Statistical methods for assessing co-evolution

Three general categories can be used to organize the computational methods used to study amino acid co-variation: (1) mutual information (MI), (2) statistical coupling analysis (SCA) and (3) directed-coupling analysis (DCA).

Mutual information. Conceptually, mutual information is grounded in probability theory and information theory. In the context of positions X and Y, MI between them, $I(X;Y)$, is a quantified expression for how much uncertainty there is at X, based on what is known about Y, or vice-versa. Mathematically, MI can be calculated for either continuous or discrete random variables, depending on the question being asked, and the latter is naturally appropriate when accounting for the set of amino acids, $\mathcal{N}$, in which each amino acid is considered to be different. At position X, we may find amino acid i with a probability of $p(i)$, and amino acid j at position Y with a probability of $p(j)$; the product of these two

marginal probabilities describes how frequently we may find i and j at these positions, if no dependency is found between them. Within X and Y, we may find the occurrence of both, amino acids i and j, simultaneously, and express this frequency as the joint probability, $p(i,j)$. Summing the ratio of the independent probabilities with this joint probability, across all possibilities of (i,j) pairs, describes the dependency between X and Y (**Eq. 4.1**).

$$I(X;Y) = \sum_{j \in \mathcal{N}} \sum_{i \in \mathcal{N}} p_{XY}(i,j) log \frac{p_{XY}(i,j)}{p_X(i) p_Y(j)} \tag{4.1}$$

A small value for $I(X;Y)$ results when $p(i,j)$ and $p(i)p(j)$ are equivalent or very similar, and thus indicates that X and Y are unrelated; on the other hand, large values for $I(X;Y)$ suggest greater disparity between the frequency of occurrences for i and j as a pair from the expected value of $p(i)p(j)$, so a stronger dependency is found between X and Y. Large $I(X;Y)$ values may occur because one particular (i,j)-pair dominates all others, or when several amino acid pairings have different joint probabilities from the expected distribution. Positive and negative values may also be adopted for every $p(i,j) log \frac{p(i,j)}{p(i)p(j)}$ term: positive values indicate that $p(i,j) > p(i)p(j)$ (finding the i,j pair more often than expected), while negative terms, where $p(i,j) < p(i)p(j)$, suggest that substitutions with i and j simultaneously is a rare event. MI essentially measures how amino acids are distributed at different positions, relative to a given one, and can be helpful in finding positions that are more constrained as a pair. However, these quantifications depend greatly on proper sampling for statistical reasons: a few hundred protein sequences should be evaluated, because 20 amino acids comprise the sample space of a given position.[139] On the other hand, MI is well-established in information theory, and can be formally extended to address larger networks of positions easily. By treating each amino acid as unique, however, biochemical features of different substitutions are generally ignored or overemphasized: large MI values arising from similar side chain substitutions may suggest a bias towards specific categories of amino acids, while large MI values due to highly disparate side-chain pairings could indicate mutational insensitivity, even as a coupled pair.

MI-based values are compared in direct-coupling analysis (DCA) to determine co-dependency between any two amino acids. For a given protein family, a multiple sequence alignment is performed and an MI score can be computed for each unique pair. The goal of DCA is to distinguish between coupling that arises from evolutionary relationships (direct coupling) from coupled interactions that can happen for other reasons, such as spatial orientation alone (indirect coupling.)[45, 44] Rather than focus on a local sequence alignment, having a global alignment of the protein family provides a better distribution of how different positions are related. MI scores for each pair is compared to all other pairs in the alignment, and statistically significant ones are believed to have evolved interdependency. The success of DCA has been used to identify important protein fold contacts for different families to predict tertiary structure from amino-acid sequences.[46] The sample size for DCA is required to be very large, in order to get statistically meaningful results, and this can be difficult to curate for some protein families. The length in primary structure can be highly variable, for instance, or not enough sequences are available to construct a reliable alignment.

Another approach that relies on multiple sequence alignments is statistical coupling analysis (SCA).[40, 41] Pairwise interactions are evaluated based on amino acid distributions between two positions: a subset of sequences are chosen so that full amino-acid conservation is found at one position, i, and the amino-acid distribution found at the other position, j, is compared to the conservation found in the original full alignment. Substantial differences between the two amino acid distributions for j are an indication that i and j are coupled; otherwise a similar level of conservation is expected, because no interdependency is constraining position j. Comparisons with an entropy-based metric is not needed here, but statistical coupling analysis can have the same issues as DCA in developing an appropriate sequence alignment. Still, SCA has been successful in highlighting amino-acid interdependency in the context of interaction specificity and allosteric mechanisms.[42, 43]

4.1.2 Applications of co-evolution methods in protein design

Predicting co-evolving pairs in any given protein system can enhance existing methods in protein structure prediction by defining important residue contacts that cannot be predicted otherwise, such as in using homology modeling. Pairs of amino acids with high interdependency are also helpful in studying protein interaction surfaces to address questions in binding specificity with ligands and with proteins. Primarily, we focus on understanding how amino-acid co-variation enables a protein to stabily fold: nearly all amino acids are assumed to contribute to structural stability in some way, and this aspect of fitness alone is fundamental to having any protein function at all.

Amino-acid co-variation is addressed in this chapter by using multiple protein sequence spaces–single, double and triple mutants–and taking into consideration how additional mutations may continue to alter fitness. Protein structural space becomes larger with an increasing number of simultaneously mutated positions, and the DEE/A* protocol is first augmented to search the protein sequence space of positions related in tertiary structure by distance cutoffs in Euclidean space. Co-dependency between positions is expected to extend beyond pairs of positions, as in the case of allosteric interactions, and the purpose of exploring an increasing number of mutable positions is to verify whether co-dependent networks of interaction will continue to grow, and the implications of overall mutational robustness in the protein as a result. Here, the modified protocol and effects of coupled interactions are demonstrated on the β subunit in $G_i\alpha_1\beta_1\gamma_2$.

4.2 Methods

The deterministic nature of DEE/A* can generate large libraries of low-energy sequences and corresponding structures, establishing an important gateway for studying amino-acid co-variation: structural context is readily available in understanding mutational effects, avoiding reliance on solely statistical analysis or defining significant thresholds to deter-

mine whether select positions are co-dependent. A straightforward extension of the pre-established DEE/A* protocol was used to study coupled mutational effects. Pairs, triplets or larger subsets of positions can be chosen to change simultaneously, adopting any amino-acid, including the wild-type residue as needed. Unsurprisingly, this will increase the search space for DEE/A*: for m simultaneously mutated positions, there are now 20^m possible sequences to evaluate, and in a protein with n positions, $\binom{n}{m}$ unique m-tuples can be taken into consideration; the structural space related to this search will grow astronomically faster, due to the increasingly large number of rotamers that must be evaluated. A protein with 350 positions will have about 2.4×10^7 and 5.7×10^{10} sequences, if all unique pairs or triplets of positions simultaneously vary, respectively. For brevity, notation for simultaneous mutations will borrow standard abbreviations used in probability theory: (A_i, A_j, A_k) changing simultaneously can be written as $(A_i A_j A_k)$ or $(A_i \cap A_j \cap A_k)$. Asterisks are used over the position that is mutated, when the distinction needs to be made.

4.2.1 Adapting protocol to address problem size

Restricting Euclidean space. Evaluating all possible pairs or triplets in the heterotrimer would not be feasible within a reasonable amount of time; there are approximately 6.1×10^5 unique pairs and 7.1×10^6 unique triplets for a protein with 350 positions—these sizes more than double for the complete heterotrimer. A reasonable assumption, however, is that positions must be close enough to interact for any co-dependency to occur, and by applying a distance cutoff in Euclidean space to address this, the number of pair and triplet interactions for evaluation is reduced considerably. A pair in this data set, $(A_i A_j)$, was chosen for mutation if at least one atom of A_i is within 5 Å of another atom of A_j; again, this 5 Å cutoff was used, because most Van der Waals interactions are expected to occur within this distance.

Triplets, $(A_i A_j A_k)$, were chosen in a similar fashion, where at least two unique pairs from the triplet satisfy the 5 Å cutoff. For example, $(A_i A_j)$ and $(A_j A_k)$ each satisfy this

distance constraint, but (A_iA_k) might be outside of it. When this happened, (A_iA_k) was simply mutated as a pair, then included into the set of double mutants for analysis. Additional constraints were applied to triplets in this approach, both to keep the magnitude of structural space manageable, and to directly address specific questions about structural stability: triplets were taken from the β-subunit only, so the interactions found are limited to understanding fitness in β-propeller proteins, and (relatedly) the requirements of β-sheets. By constraining triplets to Gβ positions, and not to the $\beta\gamma$-heterodimer or complete heterotrimer, a more detailed analysis on WD40 repeat proteins (a class of β-propellers) was also performed, and details are found in **Chapter 3**.

4.2.2 Quantifying epistasis

The DEE/A* protocol for large-scale mutagenesis provides a major advantage in evaluating protein fitness and epistatic effects. By allowing two or more amino acids to be substituted at the same time, mutable positions can be imagined as either the most recent evolutionary change to a sequence or thought of as background mutations. To understand how these background mutations may contribute to epistatic effects, the additivity between positions corresponding to a given (A_iA_j)-pair or $(A_iA_jA_k)$-triplet (ε_{ij} and ε_{ijk}, respectively) was compared: it is expected that the sum of two mutually exclusive mutations will have a nearly additive energetic effect, compared to any coupled position. (**Eq. 4.2** & **Eq. 4.3**.)

$$\varepsilon_{ij} = \Delta\Delta G(A_i^*A_j^*) - \left[\Delta\Delta G(A_i^*) + \Delta\Delta G(A_j^*)\right] \tag{4.2}$$

$$\begin{aligned}\varepsilon_{ijk} = {} & [\Delta\Delta G(A_i^*) + \Delta\Delta G(A_j^*) + \Delta\Delta G(A_k^*)] \\ & - [\Delta\Delta G(A_i^*A_j^*) + \Delta\Delta G(A_j^*A_k^*) + \Delta\Delta G(A_i^*A_k^*)] \\ & + \Delta\Delta G(A_i^*A_j^*A_k^*)\end{aligned} \tag{4.3}$$

Correcting for side-chain flexibility. Wild-type side chains may re-position into a more favorable configuration, as the number of simultaneously mutable positions increase. Naturally, this can yield lower energies that are not due to amino-acid substitution, and to account for this, a flexibility term, ε_{flex}, was introduced as an approximation (**Eq. 4.4** & **4.7**.)

$$\begin{aligned}
\varepsilon_{ij}^{flex} &= \frac{1}{2}(\sum_{u=1}^{20}[\Delta\Delta G(A_{i,u}^{*}A_j) - G(A_{i,u})] \\
&\quad + \sum_{v=1}^{20}[\Delta\Delta G(A_iA_{j,v}^{*}) - G(A_{j,v}^{*})]) \qquad (4.4) \\
\varepsilon_{1,ijk}^{flex} &= \frac{1}{3}(\sum_{u=1}^{20}[\Delta\Delta G(A_{i,u}^{*}A_jA_k) - \Delta\Delta G(A_{i,u}^{*})] \\
&\quad + \sum_{v=1}^{20}[\Delta\Delta G(A_iA_{j,v}^{*}A_k) - \Delta\Delta G(A_{j,v}^{*})] \\
&\quad + \sum_{w=1}^{20}[\Delta\Delta G(A_iA_jA_{k,w}^{*}) - \Delta\Delta G(A_{k,w}^{*})]) \qquad (4.5) \\
\varepsilon_{2,ijk}^{flex} &= \frac{1}{3}(\varepsilon_{ij}^{flex} + \varepsilon_{jk}^{flex} + \varepsilon_{ik}^{flex}) \qquad (4.6) \\
\varepsilon_{ijk}^{flex} &= \frac{1}{2}(\varepsilon_{1,ijk}^{flex} + \varepsilon_{2,ijk}^{flex}) \qquad (4.7)
\end{aligned}$$

Essentially, the difference between corresponding single-mutant sequences is computed under different mutational contexts (singles to understand doubles, doubles and also singles to understand triples.) A hypothetical example with positions ($A_1A_2A_3$) is shown to illustrate the flexible regions that are being subtracted and averaged over in **Figure 4.1**.

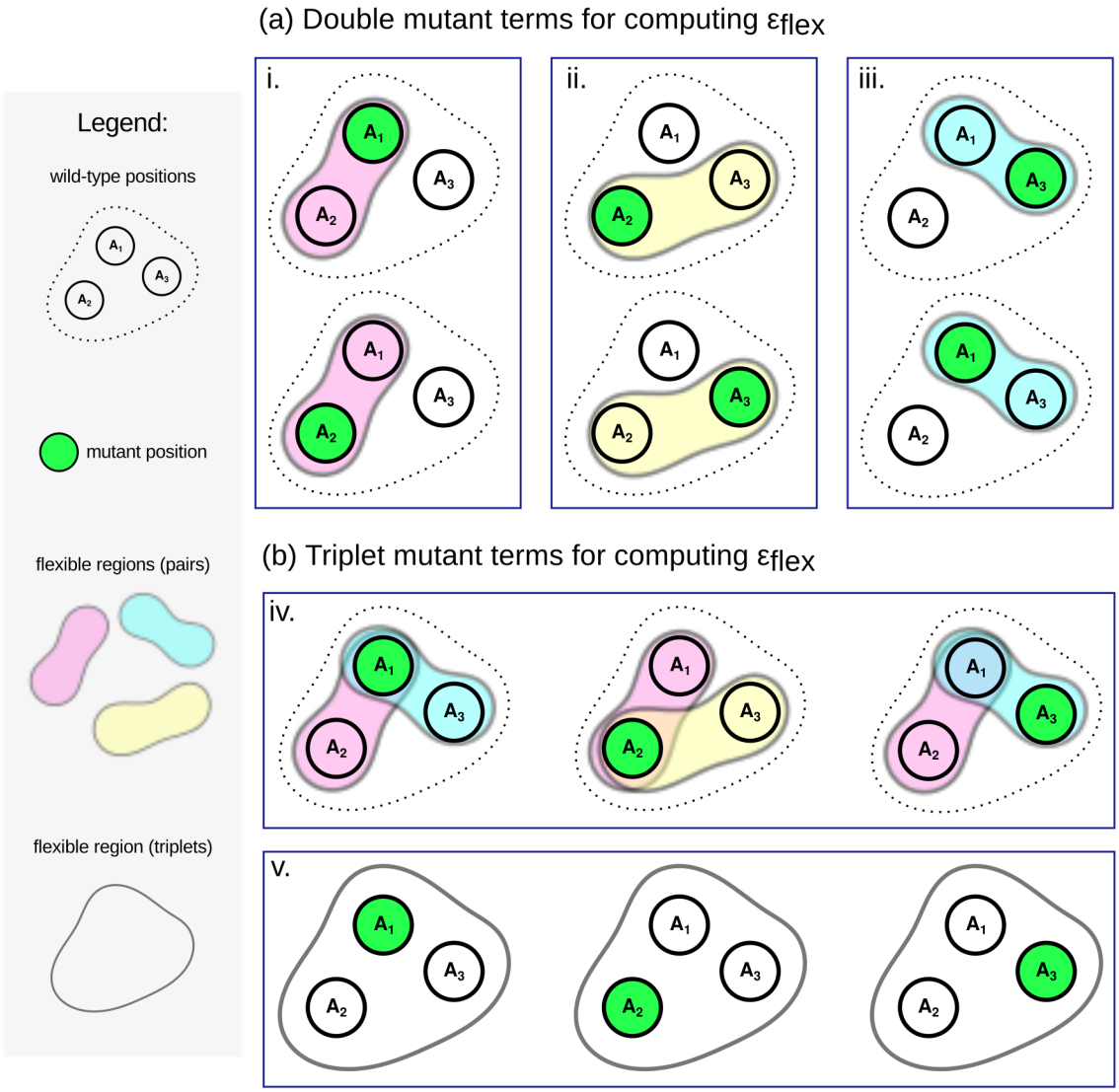


Figure 4.1: Mutable positions will sometimes prefer the wild-type amino acid, and the number of flexible side chains can be different between single, pair and double mutations (**Eq. 4.4** & **4.7**). If a wild-type amino acid is found at two positions in a situation where three wild-type positions were allowed to change, the primary structure is the same as a single mutant sequence, for instance. The average energetic variation due to a greater extent of side-chain flexibility in the protein can be computed. The example shown here is for the mutation of $(A_1A_2A_3)$. The sequences needed in double and triple mutant sequence space are shown in panels (a) and (b), respectively. A similar logic is also used to identify relevant single mutant sequences that correspond to a double mutant with only one substitution made.

4.3 Results

4.3.1 Modifications for searching larger conformational space

Directed mutations are necessary for sampling sufficiency. Unlike mutations in which only a single amino acid is ever replaced, sampling over coupled mutations (as pairs, triplets or larger subsets) presents additional challenges. Backbone conformations may be perturbed so that a pair of side chains may not be consistently close enough to interact within the context of different conformations, but the size of protein structural space for the DEE/A* search becomes larger as the number of allowed substitutions increases. Sampling a specific mutation over an appropriate ensemble of states will have trade-offs with computational expense: mutations that involve changing several positions at a time often requires more computational time for DEE/A* to successfully prune and find low-energy conformations. Using 100 nodes on the computing cluster (see **Chapter 2**), a complete set of interacting pairs on the $\beta\gamma$-heterodimer from one wild-type conformation may require 3–4 weeks of computational time, even after applying distance constraints in choosing wild-type pairs for mutation; in contrast, all single-site mutations may be completed in about 1–2 days for the same heterodimer. Amendments were made to the present computational protocol as a compromise to fully account for structural plasticity.

A global approach in which multiple snapshots are used is desirable, but will be highly varied in the unique pairs that are explored: about 40% of the data does not reappear between two different snapshots, for example, so a more directed approach in computational mutagenesis was necessary to improve consistency of sampling across different snapshots. Pairs were determined using the 5-Å cutoff from the same 40 wild-type conformations used for single-mutant analysis (**Chapter 2**) and initially compared to show variability in sampling. A total of 1795 unique pairs were identified between all 40 backbone conformations, but only about 60% of the pairs found in any individual snapshot would ever overlap with this superset (**Fig. 4.2**); naturally, pairs that overlapped with the superset were also not identical in every

snapshot. A reasonable compromise may be to include pairs that remain common among several snapshots, but shared by fewer than 40 snapshots. However, very few wild-type pairs would meet this requirement: fewer than 100 additional pairs would be incorporated into the data set, a modest change in the total sample size (**Fig. 4.3**).

In contrast to using conformations from a 350-ns simulation, it was expected that a greater number of wild-type pairs will be shared between conformations spaced closer together. The same analysis was applied to each 5-ns interval (**Fig. 4.4**), and while a greater number of pairs are sampled at the maximum frequency (nearly 80% of all unique pairs for a given interval, **Table 4.1** & **Fig. 4.5**), this is not necessarily the best approach to reconcile the sampling problem at hand. Structural conformations are not expected to vary much from one nanosecond to the next, and thus variations in backbone conformations may not adequately capture enough general biophysical features—structural dynamics are expected to happen over a longer time scale than this. Additionally, sampling backbone conformations that are timed close together is expected to yield greater conservation of usable, highly-sampled data: a larger number of pairs will be consistently evaluated in all five conformations when only a small subset of the 40 conformations are used. This same reasoning explains how a larger proportion of highly sampled pairs was found in each 5-ns interval. By defining a smaller, new superset over each 5-ns interval, only about 1300 unique pairs are found, nearly a 30% reduction from the almost 1800 possible pairs over the 350-ns interval.

Selecting a small number of conformations that span a broad range of simulation time seemed to be the most practical for balancing backbone structural variation with the larger structural space that needed to be searched: a total of five wild-type conformations, spaced 50-ns apart was used for mutagenesis. About 700 pairs (theoretically, up to 280,000 possible sequences) were mutated consistently in all five conformations, representing about 70% of all possible unique pairs (**Fig. 4.6**.)

Arguably, despite having more possibilities to choose from (20^n possible sequences

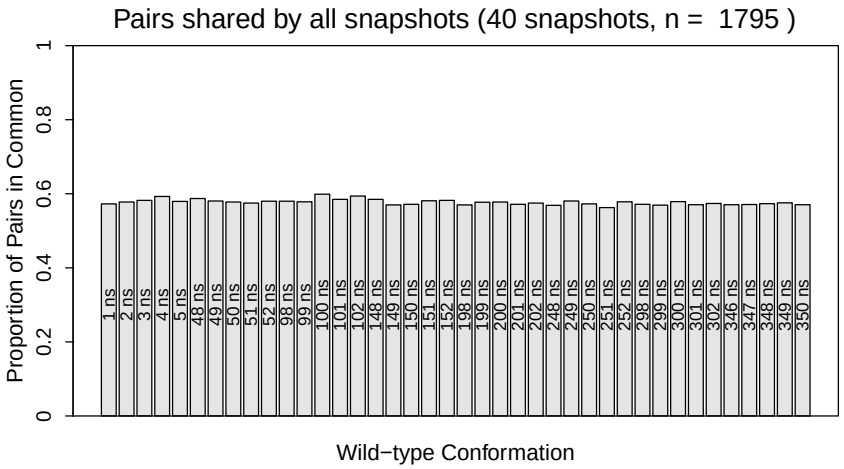


Figure 4.2: Between all 40 snapshots, a total of 1795 unique pairs are found. the proportion of pairs shared between individual snapshots to is compared and shown here; the time-point of each wild-type conformation is labeled on each bar.

Table 4.1: Conformations are listed according to time-point in the molecular dynamics simulation. The number of unique pairs found in each is also indicated.

Interval	Unique pairs
A: [1,5]	1331
B: [48,52]	1298
C: [98,102]	1334
D: [148,152]	1294
E: [198,202]	1310
F: [248,252]	1278
G: [298,302]	1288
H: [346,350]	1301

for n simultaneously mutated positions) a significant proportion of sequence space is not expected to have any corresponding structures that can satisfy the energetic criteria for structural stability defined earlier. From the single mutant sequences, it was already found that nearly two-thirds of the sequences are less favorable than wild type. Deliberately sampling specific pairs or triplets of mutation, based on previously known interactions, would significantly reduce the problem size.

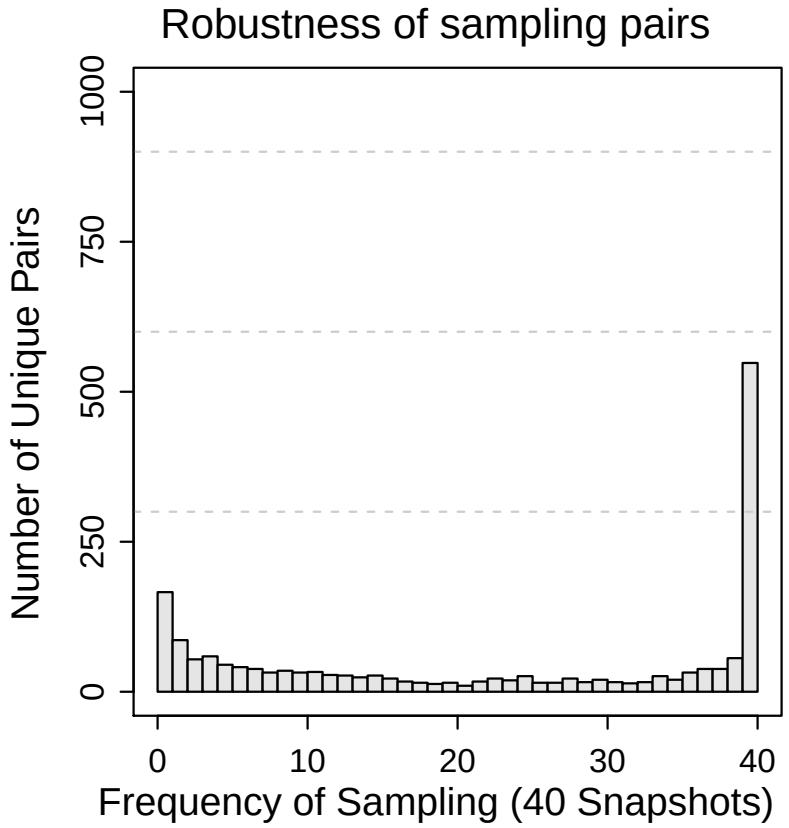


Figure 4.3: For each unique pair, the frequency of occurrence over 40 different conformations was enumerated and is shown here. Most pairs were found in all 40 wild-type conformations. Nearly 100 pairs were found only once, and another 100 are found at high frequency (in 30 snapshots or more).

4.3.2 Role of coupled interactions in mutational robustness

Coupled interactions in double mutants. The energetic differences between all single-mutant sequences within the β subunit have previously indicated that the wild-type β-propeller is very robust, and will not tolerate most side-chain substitutions (**Chapter 3**). Paired positions, based on a 5-Å distance cutoff also showed important energetic variations: mutations were generally non-additive, suggesting that coupled interactions may have an important role. At first glance, this is unsurprising, because all side-chains within the vicinity of each other are expected to have co-dependent interactions that contribute towards maintaining protein fitness. In the case of the β-propeller protein, however, the interactions define an important pattern within and between each propeller blade, with some occurring

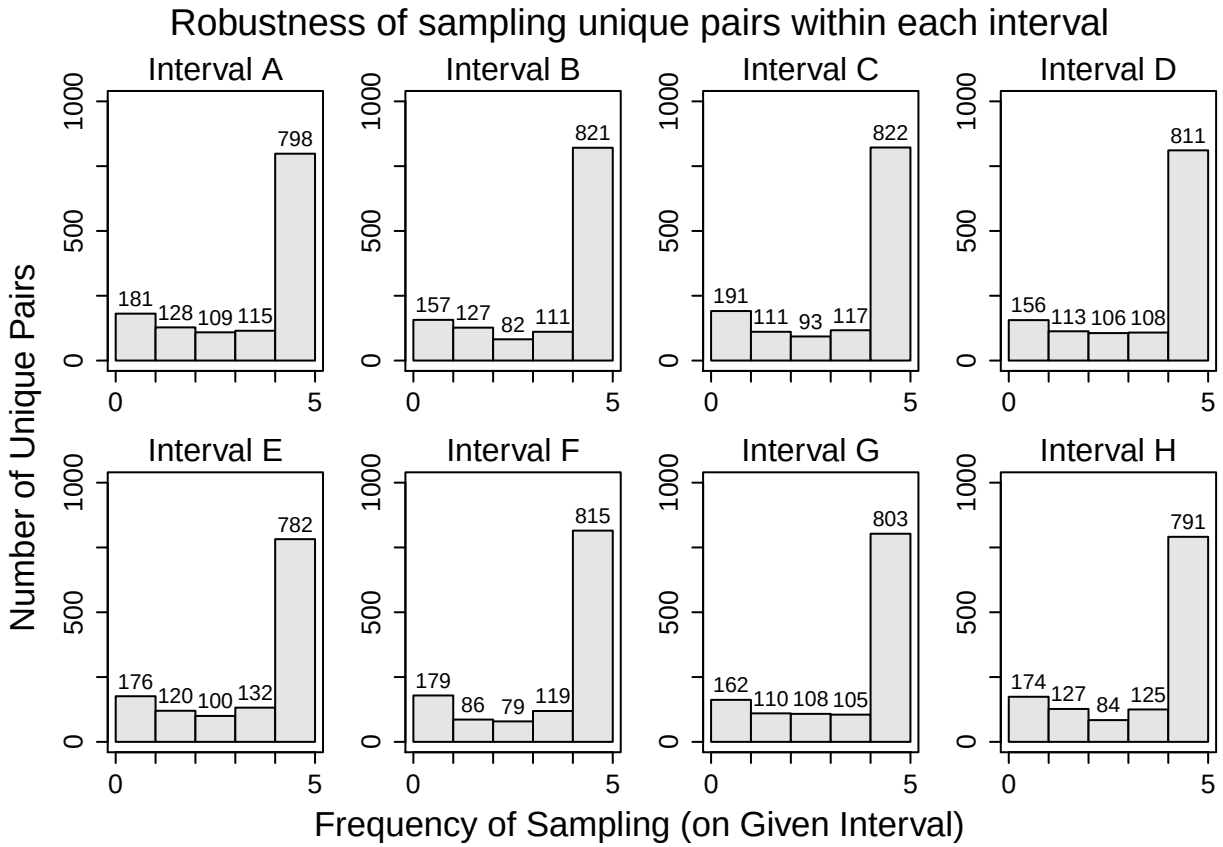


Figure 4.4: Structural invariance should yield a greater proportion of overlapping pairs. To evaluate this, the frequency of unique pairs within a given interval was counted, and are shown here. About 80% of the pairs found in any given interval can be sampled in all 5 conformations.The number of possible unique pairs is shown for each interval in **Table 4.1**.

more often than others. Here, pairwise coupled effects were explored further to see how structural integrity is maintained in Gβ.

When two positions, i and j are allowed to simultaneously vary, the mutant amino-acid at i will depend on j, and this relationship should be symmetric: substitutions at j will also depend on what is chosen at i. It may be possible that mutations occur at different frequencies when two substitutions are allowed instead of one, so the observed frequencies of finding each type of amino acid in double- and single-mutant sequence spaces were first evaluated (**Fig. 4.7**. Mutant sequences that were taken into consideration come from the β-propeller protein, and were mutated at positions i and j; for this enumeration, these sequences should have an energetically neutral or favorable result relative to wild type

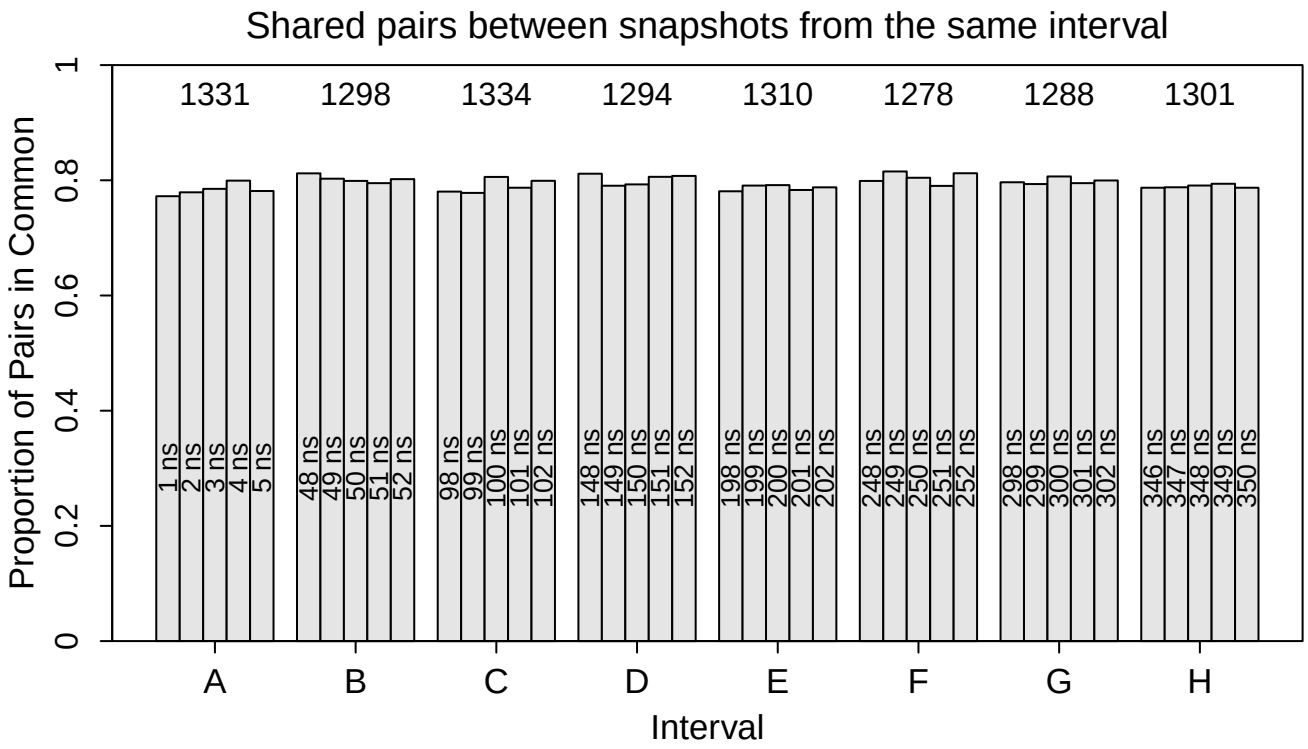


Figure 4.5: Local properties can be captured by taking sequential snapshots that are close together, but complete representative all pair-wise interactions is still not possible. About 80% of the pairs in a given interval are consistently found in all other conformations of the same interval.

after mutation, but only in terms of structural stability, $\Delta\Delta G_{fold}$ (binding interactions were generally negligible.) Double mutants were counted twice, once for the change at i and again at j. A total of 29,490 unique double mutants could meet this criteria, while 2652 unique single mutants were found. First, the probability of finding each wild-type–mutant amino-acid pairing was determined, and the distribution from each energetically constrained protein sequence space was compared. In total, there were 324 points for each distribution (18×18, since Pro and Gly are excluded), and the frequencies of substitution were similar. Although the Pearson's correlation coefficient, ρ^2, is almost 0 based on a linear fit (red line), some amino-acid pairings were never found as double mutants due to additional biophysical constraints; this yields a string of 0 probabilities which weakens overall correlation by this metric. However, most of the data lie on or near the best-fit line, $y = x$ (gray), which

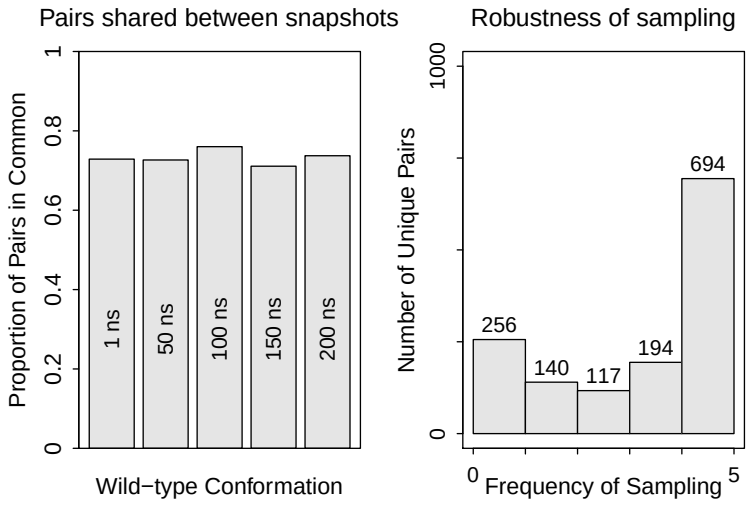


Figure 4.6: The proportion of found pairs from each snapshot is shown on the left panel. About 70% of the unique pairs in each overlapped with the overall superset from all five conformations. The frequency of sampling each unique pair in the superset is illustrated on the right. A few hundred pairs were found in two, three or four of the chosen snapshots, and most pairs were found in all of them.

suggested that some correlation between the two data sets. Still, this relationship was a reflection that the choice of amino-acid in replacing wild type would need to vary when substitutions were allowed to simultaneously change.

The types of substitutions that are found can also be assessed, by counting how often a specific mutant amino acid appears in neutral and favorable regions of protein sequence spaces (**Fig. 4.8**). The probability of finding a specific amino acid in single-mutant sequence space was proportional to what was found in the double-mutant sequence space. This demonstrates that while specific amino-acid replacements might be avoided in double mutants, the general biophysical features that lead to specific amino acids being preferred over others could still be found.

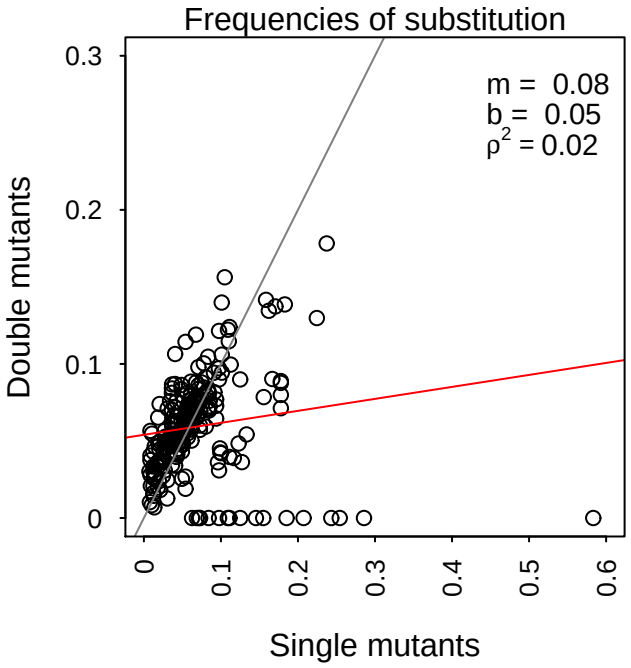


Figure 4.7: The frequency of amino-acid substitution from a given wild-type amino acid to an alternative amino acid was enumerated for all sequences $\leq$ 1.5 kcal/mol in structural stability, relative to the wild-type sequence. These substitution rates were based on the 5 conformations used to evaluate pair-wise interactions in the β-propeller; single mutants from the same 5 conformations were used for comparison. There are 29,500 double mutants considered here, Boltzmann-weighted over the 5 different conformations. Substitutions in single-mutant sequences follow a relatively uniform distribution, and variance in these values were typically low. In considering interactions between two mutant amino acids, the distribution was more highly varied, suggesting strong biophysical constraints were imposed when two modifications need to be made to the wild-type sequence. The best-fit line to this data is shown in red, with the slope and intercept are indicated within the correlation plot; for visual reference, $y = x$ is also shown (in gray.)

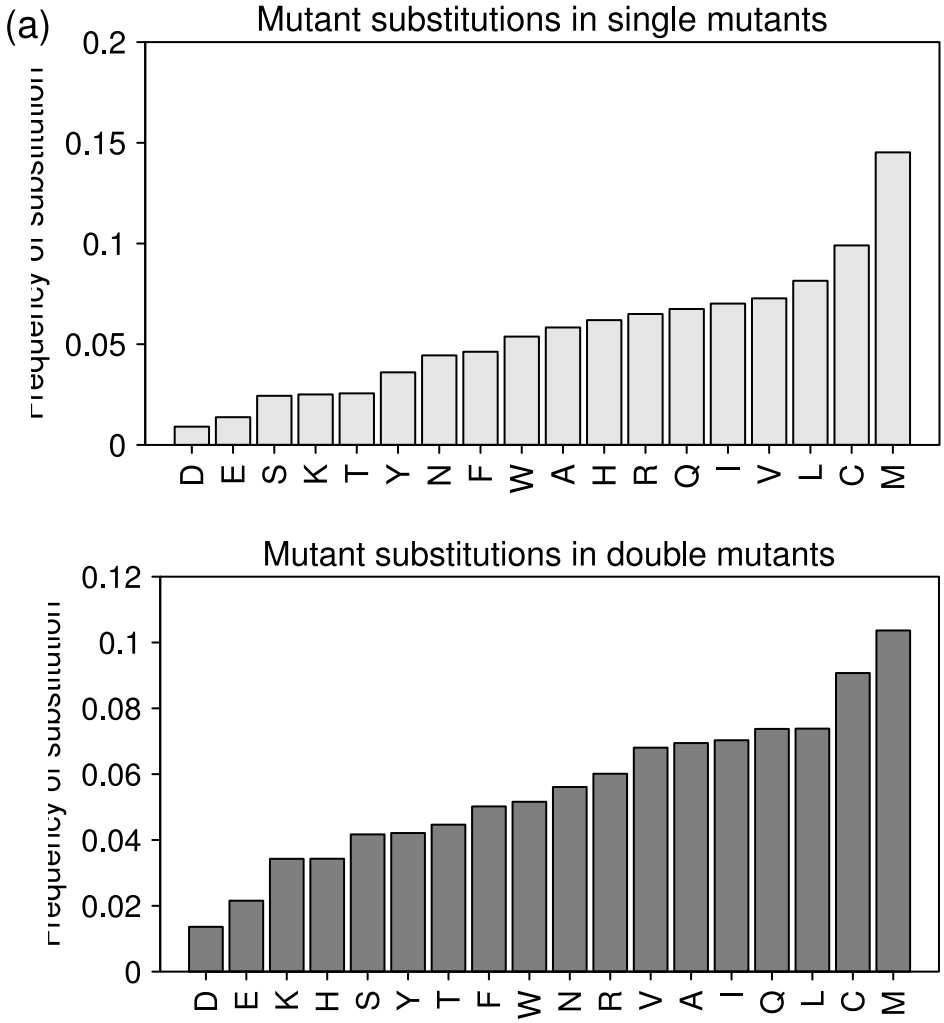


Figure 4.8: Sequences that were energetically favorable or neutral relative to wild type were enumerated. All sequences considered here are fully mutated (must have one mutation in (a) and two mutations in (b). The frequency of finding substitutions to each amino acid is shown here as a probability based on all mutant substitutions. A shift in the distribution is a consequence of having more stringent requirements when two positions are simultaneously changed instead of one only.

Energetic contributions of nearby wild-type residues. The effect of having a greater number of wild-type side-chains adopt alternative configurations in the protein, ε_{flex} was taken into consideration, for double and triple mutants. Depending on the proximity of the positions that were allowed to mutate, the number of re-orienting wild-type amino acids could be modest if the mutable positions were rather close. There were 266 unique triplets evaluated, and the distribution of ε_{flex} for a fully mutated triplet was typically less than 10 kcal/mol (**Fig. 4.9**). The sampling size was small, considering the range of ε_{flex} values could almost be 60 kcal/mol, so the kernel density estimate was provided and suggested that large (>20 kcal/mol) energetic changes from wild-type re-orientation had a very low probability of occurrence. At the same time, the probability was also unlikely when ε_{flex} was approximately 0 kcal/mol by this estimate. In most cases, the energetic change due to surrounding wild-type amino acids was significant, and thus should be accounted for when evaluating double and triple mutants.

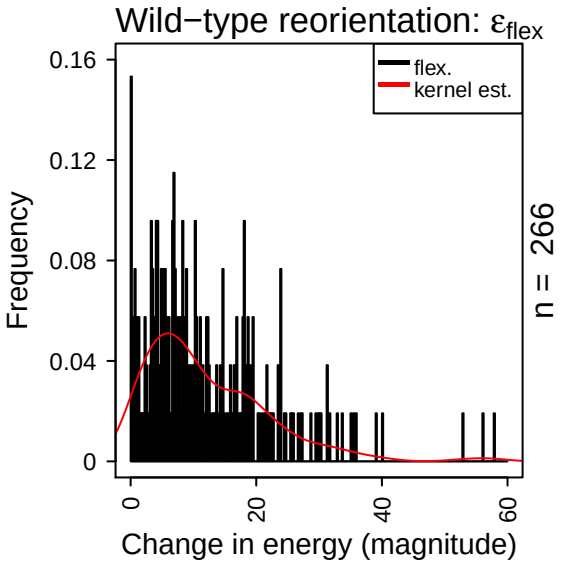


Figure 4.9: Attempting to mutate 2 or 3 positions simultaneously can sometimes yield sequences in which one or more of the positions remain as the wild-type residue. The number of neighboring side-chains may also increase, as the region of mutable positions becomes larger. To account for this, ε_{flex} energetic change in a sequence due to configurational changes in the wild-type side-chains, can be computed from appropriate corresponding sequences that make up mutated positions.

Observed epistasis between double and triple mutants. From the set of triple mutants, corresponding double and single mutants were used for comparison to understand how ε_{epi}, the magnitude of epistasis between relevant triple, pair and single mutations, could vary (**Fig. 4.10**). Most pairwise interactions yield a non-zero epistasis value, even after correcting for wild-type flexibility (**Fig. 4.10a**). Values that were essentially zero may be due to numerical errors. Regardless, very few double mutants exhibit this, and a greater tendency to have most ε_{epi} values on the range of 10 and 30 kcal/mol was found. Over 33,000 unique sequences were used in this assessment, so the distribution of real data (black) and estimated distribution (red) overlap almost perfectly. Energetic data for single and double mutants were shuffled to determine the extent that epistasis was truly sequence dependent, rather than arbitrarily restricted; the new distribution of ε_{epi} was computed after randomization, and 1000 replicates were performed (**Fig.4.10**, shown in gray). Obviously, the two distributions had very different behavior: when sequences were completely shuffled, the fabricated epistasis effect was very large, falling onto a distribution that was more broadly distributed than what was found in the actual double-mutant sequences and this verified that ε_{epi} is not artifactual.

More than 44,000 triple mutants were also used in the same type of analysis (**Fig. 4.10b**), and a broader distribution in the actual data resulted, relative to what was found among the double mutants. Larger epistatic effects overall are not completely surprising: more coordination and co-dependency should be expected between three different positions that need to interact with each other than between two of them. At the same time, though, the amount of sampling required to sufficiently represent overall epistasis in triple mutations may be larger than the number of sequences that were used in the analysis. For example, from the 266 unique triplets of position, there are more than 2.1×10^7 possible triple mutants—only 2% of these sequences were considered low-energy, found by DEE/A*.

With this in mind, that the magnitude of epistasis may be highly variable in some cases, but generally a significant value among double and also triple mutants, the average

behavior of an increasing number of simultaneously changing amino-acids was assessed (**Fig. 4.11**). All sequences were first corrected for the contribution of ε_{flex} to provide a better estimate of how the mutation itself has an effect, instead of the wild-type side chains that were not allowed to change amino acids. Triplets were defined according to the frequency of occurrence in the pairwise motifs (low, if found in 4 or fewer blades, and high if found in 5 or more), regardless of whether inter-repeat or intra-repeat interactions were made, and also by approximate location on the propeller blade (see **Chapter 3**). Additionally, triplets and pairs were assessed for the DHSW motif as a frame of reference to determine how coupled interactions behave in a well-known structural motif that has been demonstrated to be mutationally robust.[129] Five general categories were used to separate all triple mutants for analysis (**Fig. 4.11**), based on structural conservation: (i) DHSW motif, (ii) coupled interactions within the β-subunit tunnel (based on strand a positions), (iii) triplets that involve only positions from either strand b or c (high frequency as pairs), (iv) additional triplets that involve only strands b and c (low frequency as pairs), and finally (v) inter-repeat triplets between adjacent β-sheets (strands a with b, strands b with c, and c with d). The expectation was that structural conservation and greater mutational robustness should be found in (i), (ii) and (iii), while relatively fewer structural requirements were expected in groups (iv) and (v).

For each mutant triplet (A_i^*, A_j^*, A_k^*), the average effect of a single mutation (i.e. $\frac{1}{3}\left[\Delta\Delta G_{fold}(A_i^*) + G_{fold}(A_j^*) + G_{fold}(A_k^*)\right]$) and the average effect of a double mutation were computed (i.e. $\frac{1}{3}\left[\Delta\Delta G_{fold}(A_i^*, A_j^*) + \Delta\Delta G_{fold}(A_j^*, A_k^*) + \Delta\Delta G_{fold}(A_i^*, A_k^*)\right]$), using the appropriate sequences. The energetic distribution of these terms for all triple mutants were used as an assessment for how the accumulation of new mutations could change the structural stability of a given triple mutant. The effect of mutations from single to double to triple mutant could also be considered by taking the energetic difference between corresponding sequences (i.e. $\Delta\Delta G_{fold}(A_i^*, A_j^*) - \Delta\Delta G_{fold}(A_i^*)$ and $\Delta\Delta G_{fold}(A_i^*, A_j^*) - \Delta\Delta G_{fold}(A_j^*)$). Interpreting the sign between these differences can become complicated, however; the approach

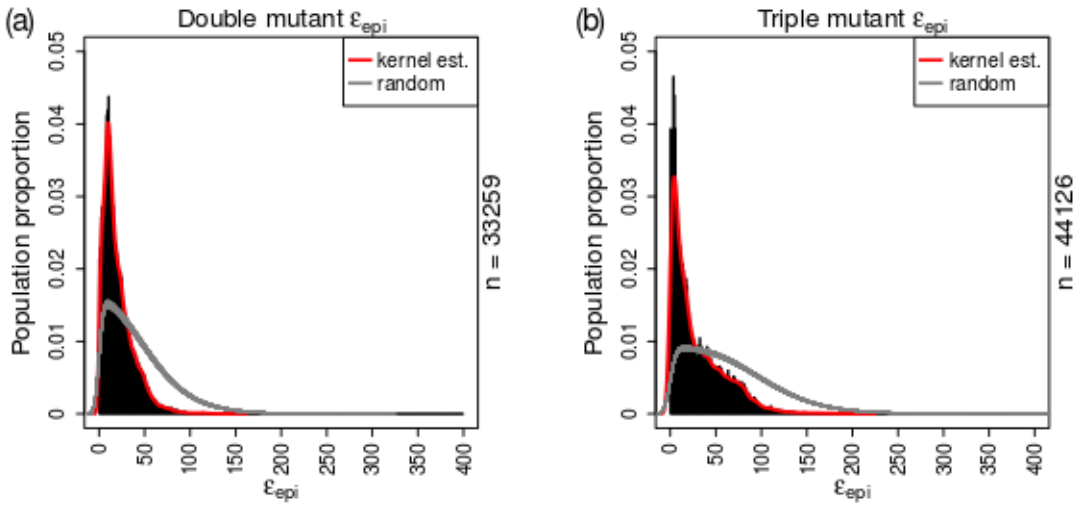


Figure 4.10: Here, the magnitude of epistasis from having (a) two mutations or (b) three mutations are shown here in black. The population of double mutants were the corresponding unique pairs that were also used to define triple-mutant positions. The kernel for this distribution is shown in red. For comparison, energetic data was also randomized, prior to calculating ε_{epi}; in each (a) and (b), 1000 such random samples were created, and the kernel density estimate of each is shown in gray.

used here instead focuses more on the resulting fitness from each kind of mutation, and it becomes obvious that most sequences are less favorable than wild type. Moreover, the energy distribution shifted towards being more unfavorable when additional positions can be changed simultaneously, and this ultimately supported that the β-propeller considered here is a highly evolved protein with important coupled interactions in the wild-type structure: attempting to find a new combination of amino acids to fit within the pre-existing biophysical framework should be challenging, because the number of co-dependent interactions is greater and potentially stronger.

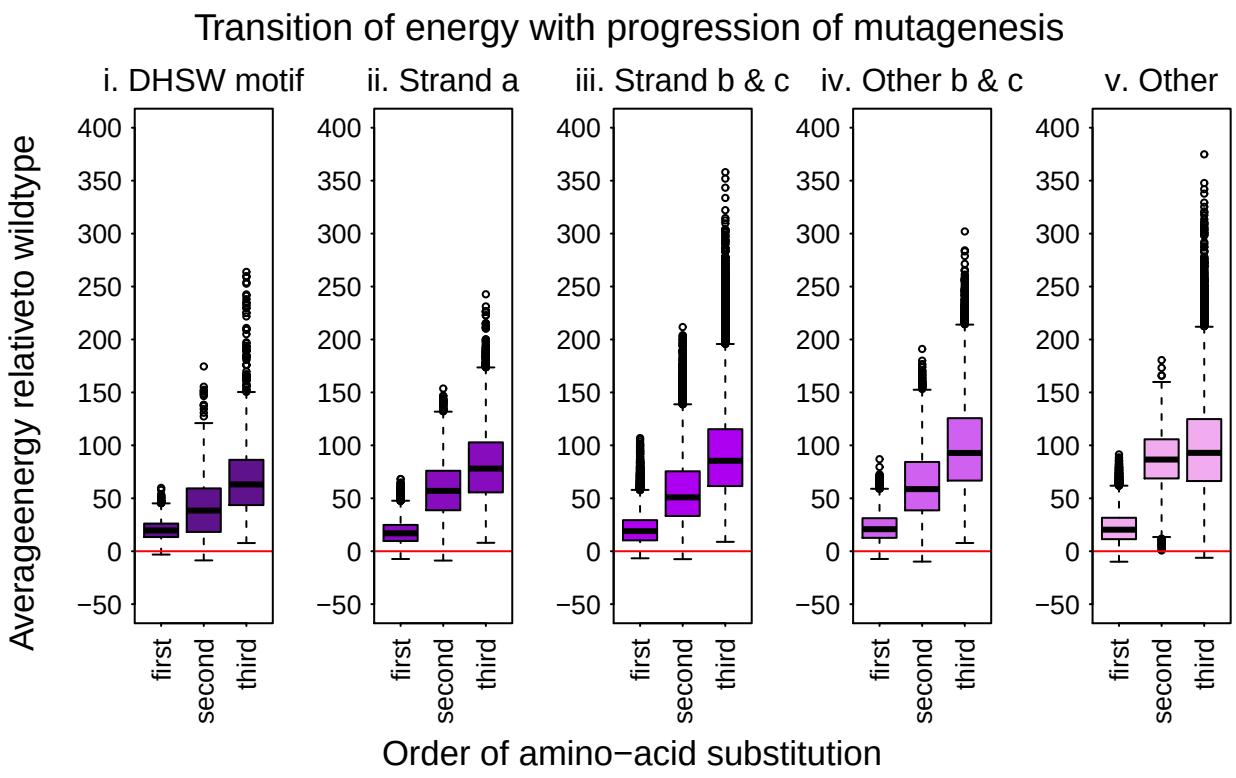


Figure 4.11: Average energy distribution of all triple mutants, and corresponding double and single mutants are shown here (i.e. $\langle \Delta\Delta G(A_i^*, A_j^*, A_k^*) \rangle$, $\langle \Delta\Delta G(A_i^*, A_j^*), \Delta\Delta G(A_j^*, A_k^*), \Delta\Delta G(A_i^*, A_k^*) \rangle$, and $\langle \Delta\Delta G(A_i^*), \Delta\Delta G(A_j^*), \Delta\Delta G(A_k^*) \rangle$.) The intensity of purple is loosely related to mutational robustness expected for the motif, based on structure: (i) DHSW motifs and (ii) the β-propeller tunnel (comprised of strand a positions) are expected to be very sensitive to change, while (iii) the two β-sheets within a repeat (strands c & d) is moderately sensitive to changes. Groups (ii) & (iii) are also highly connected in the interactions mentioned in **Chapter 3**. Groups (iv) and (v) are triplets from pairs that are not frequently found to make coupled interactions.

4.4 Conclusions & Discussion

Accounting for coupled interactions is an important aspect of assessing sequence–function relationships. Whether using analytical methods that are statistically based or, as demonstrated here, in simultaneously exploring protein sequence and structural spaces, a major obstacle to fully understanding the interactions underlying protein fitness requires knowing how subsets of a larger mutation need to interact: triple mutations need to be assessed using double and also single mutants, for instance, to address this. Having enough sequences within each of these subsets, though, can provide a major advantage in constructing hypo-

thetical mutational pathways that may potentially illustrate how fitness changes with the accumulation of mutations in a given protein sequence. By extending the DEE/A* protocol to find these sequences, a number of important observations were made about β-propeller proteins.

Initially, single-mutant sequences were generally less favorable than wild type (**Chapter 2**), and coupled effects were also found to be less structurally stable than wild type (**Chapter 3**). Here, the progressive changes were identified from one set of mutations to the next, and as the number of simultaneously mutated positions increased, the corresponding energy also shifted and became more unfavorable relative to wild type. In addition to this change, the network of mutationally robust interactions was extended, and more tightly knit than what was previously found. Many of the unique pairs could have a shared position and continue to be very strongly coupled. Some biases for this result do exist: many triple mutations were based on positions within strands c and d, so it was less obvious if highly coupled interactions could be found on a larger scale, if mutations limited to positions on the outer-most β strands were evaluated instead. A mixture of pairs were used to construct these triplets, based on different frequencies of coupled interactions found on the seven propeller blades; if triplets were based on pairs that have no previous observed signal (defined as $\varepsilon_{epi} \geq 0.5$ kcal/mol in **Chapter 3**), the possibility of finding sequences better than wild type may be slightly better.

On the other hand, it's important to remember that the β-propeller protein is not an ideal candidate for finding substitutions that are better than wild type. Being a highly evolved and extensive protein family, it is expected that evolution has found most of the key interactions that are necessary for keeping the repeating regions together, and the protein very stable and functional. For a computational algorithm to potentially map out these interactions, though, is also interesting: this protein has seven-fold symmetry, which (presumably) can be exploited rather easily. Each repeat is relatively small in size for both primary and tertiary structures. Then, it may be possible to limit the extensive search of

sequence and conformational spaces to only two or three of the propeller blades, and still find a more complete representation of all major coupled interactions in the protein. However, the sampling used in the current analysis would need to be improved first: pairs and triplets chosen for mutations should be picked based on a fewer criteria, and the approximations used to gauge the effect of wild-type side-chain re-orientation should be computed more explicitly. With these additional amendments, a well-defined framework of interactions could be determined and used for different purposes, from understanding interaction specificity within each subtype of β subunits (since the contributors to stability and binding can be determined) to decomposing the interaction network to identify the most basic set of coupled residues required for stability (providing a possible canvas for designing functionally different proteins.)

Chapter 5

General Conclusions

5.1 Introduction

Embedded in protein primary structure are the set of instructions that lay out important molecular interactions required for a given protein to remain biologically functional: from stabily folding into the correct three-dimensional structure, to associating with the appropriate binding partners, dynamically changing to accomodate allosteric mechanisms, and more. For a given amino acid, accomodations can only be made towards satisfying all these attributes of protein fitness, and several more, if amino acids can adopt multiple roles in maintaining overall fitness. Understanding how amino acids take part in each aspect of fitness, exclusively, requires careful analysis of how substitutions to wild type influence the initial level of fitness.

Depending on how wild type is affected, mutagenesis studies can be used to evaluate the functional roles of amino-acid substitutions: replacing residues that are essential to the aspect of fitness being studied should result in decreased fitness overall, while less important ones will have little or no effect. Currently, several methods exist for mutagenesis, but are often limited to small proteins or peptides, or only focus on a specific region of protein sequence space—for example, either mutagenesis is restricted to specific positions

or domains, or the protein sequences may be predisposed to selection pressures, from natural evolutionary processes or pre-determined crtieria in directed evolution. Simultaneously searching over protein sequence and structural spaces with the dead-end elimination and A* search algorithms enabled the construction of an unbiased protein sequence space: as long as the corresponding structure of a sequence was low in energy, it was included in analysis. With far fewer constraints and biases than evolution- or sequence-based methods, a more complete assessment on protein mutability or sequence invariance could be developed: exhaustively mutating all positions to all possible amino acids (except Pro and Gly) provides a major advantage in understanding how molecular interactions are made in all situations, whether fitness is improved, unchanged or worsened.

Computational protein design algorithms can explore protein sequence spaces extensively. The heterotrimeric G-protein used throughout this study, $G_i\alpha_1\beta_1\gamma_2$, was mutated at 685 positions to understand how single mutations can influence protein fitness, defined here as a combination of structural stability and binding interactions. Theoretically, the full protein sequence space has 13,700 sequences (His can adopt 3 possible protonation states, so a different 20 amino-acid alphabet is used.) Among these, DEE/A* was able to systematically find 13,586 of them (about 99% of the possible sequence space) using about 8 weeks of computation time to evaluate all 40 protein conformations. These sequences were used to demonstrate that only one in three mutations may yield a sequence that can stabily fold with similar stability as the wild type or better, and in total about 15% of all sequences may be structurally stable and still bind with either Gα or the $\beta\gamma$-heterodimer.

All mutations were mapped out according to position in primary structure, and a strong correlation between mutational robustness and structural context was found: mutations adversely affected positions that are at the binding interface of Gα and the $\beta\gamma$-heterodimer most, and compact, buried regions with well-defined secondary structures were more difficult to improve than at loops or more solvent-exposed locations. These observations were consistent with what is well-known about protein secondary structure features.

To validate that DEE/A* can make biophysically rational choices, the types of mutation that maintained or improved both stability and binding interactions were compared to expected substitution rates derived from sequence alignments of known proteins. Although the correlation was strong, it also verified that implementing DEE/A* to understand all aspects fitness perfectly would not be possible here—several factors beyond structural stability and binding between cognate parters are involved to achieve fitness in biological systems, and these were not measured by the protocol. However, from using DEE/A*, a very complete protein sequence space and their related structures becomes accessible, and the computational resources used are considerably economical, especially when compared to the hypothetical expenses of analogous biological experiments with the same level of unbiased sampling through mutagenesis.

Stabilizing interactions follow radially symmetric patterns in β subunit. A common problem in many approaches used to determined sequence–function relationships is the limited description, if any, of amino-acid co-dependency between pairs of interacting side chains. Amendments were made to the DEE/A* protocol to address how amino-acid covariation plays a role in maintaining structural stability. The structural space that DEE/A* must search becomes significantly larger when an increasing number of positions are allowed to simultaneously vary—a greater number of rotamers would need to be evaluated by the algorithm—so pairs of positions were limited to the β subunit, and required to be within a distance cutoff from each other. A smaller number of protein backbone conformations were also used in the sequence–structure search for the same reason. From using only five backbone conformations, about 700 unique pairs were consistently evaluated in total.

Coupled interactions were found throughout the β subunit, at varying magnitudes. The structural integrity of the β subunit is well-recognized:[125, 126, 138] the seven-fold symmetry and orientation of anti-parallel sheets have been suggested to be very stabilizing features for the β-propeller family. How mutational robustness affects the entire protein, however, is less well-known. Regions of stability have been identified, and also proprosed

in theoretical analysis based on β-propeller geometry, but mutagenesis studies are rarely performed. To understand how the radial symmetry contributes to structural stability, a reference blade was constructed to map out the repetition of coupled interactions. Positions that were most frequently coupled were found at the central tunnel of the β-propeller protein, and additional important interactions were found between them and adjacent β-sheets, both within the same strand a, but in a neighboring repeat, or at strand b within the same repeat. Many of these interactions were found in all seven propeller blades, and most pairwise mutations at these pairs of positions were generally less favorable than wildtype. Although a meaningful network of interactions were formed from these coupled pairs, the symmetric nature of the protein also suggested that co-dependency may be more extensive—a single pair of positions might not be sufficient for holding two neighboring blades together, for instance, and this motivated a search for significant coupling between a greater number of positions.

Mutational robustness is more prominent with higher-order mutagenesis. Relationships between different unique pairs were explored by constructing triplets from coupled pairs of positions. Coupled interactions varied in the number of repetitions on the reference blade, and the magnitude of epistasis. Each triplet was made from two pairs that had a common position together, and over 250 triplets were evaluated. Only one protein conformation was used as a compromise to being able to explore a broader set of triple mutants.

While low-energy sequences could still be found, most remained less stable than wild type; those that were initially observed as being favorable relative to the wild-type sequence were never triple mutants—at least one position remained as the original wild-type residue, indicating that favorable energetic changes were due to side-chain re-orientation instead of a mutational effect. A gradual loss of fitness as mutations continued to accumulate could also be observed, due to the availability of single- and double-mutant sequence spaces. On average, the interactions by single mutants were sometimes stabilizing or were a modest

improvement to wild type, but the fitness became worse when more mutations were added. This change was not necessarily dramatic, and since all mutant sequences were relative to wild type, it is possible that some mutant sequences improved in fitness, relative to a previous state with fewer mutations in the context of a different reference state. More interestingly, while this behavior was expected at highly conserved structural motifs, like DHSW or the channel formed from strand a, similar patterns were found in all groups of triplets that were evaluated. One possible interpretation is that the combined coupled effect is simply stronger when a greater number of interactions are considered. On the other hand, it may also be possible that important coupled interactions naturally happen on a larger scale for this protein, but further validation is required to confirm this. Overall, these results are important in demonstrating that evolution has a lot figured out—β-propellers are a common protein fold, because the family has become highly evolved and provides a foundation for modulating many different biological processes, and it is expected improvements on the native sequence are difficult to make for this type of protein family.

Insight on amino-acid contributions to specific aspects of fitness. Although sequences were found by DEE/A* based on a combination of structural stability and binding interactions, the energy for each specific feature could be isolated for analysis. Decomposition of the energy function is a primary reason why this computational approach could successfully identify whether a mutation would have an impact on either aspect of fitness. The efficient search over sequence and structural spaces is significant, because the relationship between amino-acid sequences and three-dimensional structure can be understood. Most amino acids in primary structure are expected to contribute towards protein folding interactions, because correct protein function relies on tertiary structure; it was unsurprising to see that very few positions could be mutated to a more favorable sequence than wild type. Very few positions are found overall at the binding interface, and the DEE/A* results were able to reveal where important interactions need to be made; all other positions had little or no observable effect. Also, some trends were found in sequences that were similar to or

better than wild type: amino acids that are either charged, polar or bulky have a lesser probability of being found, compared to aliphatic residues and smaller ones. This tendency is consistent with and expected for a typical protein. Charged and polar residues favor solvent-exposed environments, for example, and most of the mutated positions are buried, so many substitutions to these amino acids should be less favorable than wild type; discriminating between amino acids in a way provides a more convincing argument that this computational approach takes biophysical interactions into account.

5.2 Areas for improving the DEE/A* protocol

Implementing DEE/A* to address protein fitness requirements via mutagenesis can be a helpful step in experimental design. The accuracy of DEE/A* has not been fine-tuned here for imitating biology, but it can still provide a very thorough analysis on where changes could ever be made in a given protein. For experiments on protein systems in which an appropriate starting point may not be clear or obvious, applying this computational approach can be especially helpful in gauging how substitutions may affect the protein of interest. In the context of designing proteins, for example, the DEE/A* protocol could help identify regions to avoid, areas to target, and general features of allowable substitutions, potentially reducing a problem to a more manageable size. For a more accurate set of computational results, there are a few aspects of this protocol that can be further improved.

Validation with biological experiments. A single change in amino acid can completely alter protein tertiary structure, and the extent that proteins become unfolded or disordered is difficult to predict. The reliability of the DEE/A* algorithm was tested using alternative computational approaches, and the strongest correlation made to biology was based on expected substitution rates from similarity matrices as well as structural studies performed by others. Together, these approaches showed that the scope and details of protein sequence space are reasonable and there is reason to be confident in the results. However, the

ideal comparison is a more direct comparison with additional *in vitro* or *in vivo* experiments. A problem with using $G_i\alpha_1\beta_1\gamma_2$ as a model system is that heterotrimeric G-proteins are very difficult to isolate and express; it would not be possible to exhaustively mutate this protein outside of using computational algorithms, and while a number of studies exist in describing heterotrimeric G-protein pathology and experiments have been performed to evaluate binding interaction specificity for this family, the results are not always comparable for this particular protein. A compromise for this may be to test specific mutations that seemed interesting from the computational data. Alternatively, finding a protein system that has been thoroughly explored with mutagenesis studies, like those from high-throughput sequencing, would also be helpful. Protein sequences from these experiments have already been screened for mutations that affect a specific type of binding interaction, for example, and because protein size tends to be smaller for these experiments, a molecular dynamics simulation may converge sooner, and a greater number of conformations from it could be explored when implementing DEE/A*.

Accuracy of energetic calculations. As mentioned in **Chapter 2**, a number of energetic models could be used in concert with DEE/A*, and having a hierarchy of them can be useful: good approximations can be made using coarser-grained energy models to filter and separate more interesting sequences for evaluation with more intensive and accurate calculations. Even without changing the energetic models that have been used in the current setup, it is possible to take advantage of the structures that are generated from the DEE/A* results. The comparison between using a rotamer library and energy minimization has already been made here, and revealed that most conformations found by DEE/A* using a modified Dunbrack–Karplus library are already in a low-energy state, and the Newton–Rhapson algorithm could not improve further. It is reasonable to use an even finer rotamer library with DEE/A* to understand mutational effects with even greater confidence, but a more rigorous approach would be to take the mutant structures of interest and perform molecular dynamics simulations on them, fully exploring related backbone and side-chain

orientations.

Depending on how distant different mutations are on the protein, it may be possible to simulate multiple mutations on a given protein all at once. If the positions are very apart, only one trajectory that involves all of them may be needed; the results on each mutant region can be analyzed separately and interpreting them should be more straightforward than with proximal positions. Exploring very close mutations with a molecular dynamics simulation, however, can also be very beneficial. While it may be harder to disentangle the contribution of each side chain towards fitness, the trade-off is in providing additional perspective on how coupled interactions are required for the protein system. The practicality of doing any molecular dynamics simulation will partly depend on the protein size: longer simulations may be required for accurate sampling of larger protein systems, for instance, but the usage of computational resources for this may still be more affordable than attempting biological experiments without a more careful assessment first.

Improving computational efficiency of algorithms. Searching over all amino-acid rotamers at multiple positions that are allowed to change simultaneously has been a bottleneck for completing triple-site mutations, and sometimes for pairwise mutations. Accommodations were made by lowering the energetic cutoff for the A* search, and all of the low-energy sequences found performed worse than wild type, providing reassurance that favorable solutions were not missed despite reduced sampling. At the same time, these solutions required several days of computing time to be found, so the results could seem unsatisfying after such a long waiting time. There were also situations when a computing node would run out of memory and could not evaluate all of the low-energy solutions. To circumvent this problem, the DEE-pruned conformational space could be divided into more manageable pieces, and the A* search algorithm may be futher adapted to evaluate them using a parallel computing approach. Each sub-division could be assigned to a different node for the A* search algorithm, but amendments would need to be made on how the heuristic function is used. The entire set of DEE-pruned rotamers will not be on the same computing

node, so evaluating the optimal path and the ones that follow it would need to be done more carefully, and communication between processing nodes that are involved in the same search problem would need to be considered. A possible solution to this is for each node to evaluate the sub-division first, then merge the solutions from all processors that were given a part of this task, and do a second evaluation based on the estimated values from the first.

5.3 Suggestions for future directions

Throughout this thesis, the protein sequences and corresponding structures found by the DEE/A* algorithms have been used to address how protein fitness can be deconvolved, so that the contributions of each amino acid in a heterotrimeric G-protein can be understood. The efficiency of DEE/A* and scale of mutagenesis that can be performed by it is rather exceptional, and a natural continuation would be to tackle related problems that would also benefit from having a good approximation of a protein fitness landscape. In particular, a few areas in protein evolution may be interesting to investigate.

Neutral mutations and network connectivity. Neutral mutations have been known to provide some leeway for proteins to have otherwise deleterious mutations become fixed into their protein sequences. In many ways, neutral changes open up inaccessible areas of protein sequence space, and within this restricted area, there may be protein sequences that can have improved stability and function as a consequence of naturally random mutations. In neutral evolution theory, a widely accepted assumption is that most mutations will have a neutral effect, and how the accumulation of neutral changes contribute to mutational robustness for a given protein is not well understood. A common approach is based on graph theoretical methods to illustrate the relationship between neutral mutations and more progressive ones: unique sequences are represented by nodes, and edges are formed between any pair of them if there is a pathway to change one into the next. With this representation, highly connected nodes are thought to have more options for neutral evolution,

and remain mutationally robust, while nodes with less connectivity have the ability to evolve relatively sooner in a way that would improve or worsen fitness. Network modeling provides many benefits to visualization, and can also follow many mathematical principles, which can facilitate analysis and interpretation of sequence space.

Many theoretical analyses, however, have used very simple two-dimensional models, such as hydrophobic–polar lattice models to demonstrate the results of neutral-mutation accumulation. An amino-acid sequence is often short (about 20 amino acids) and is represented by beads that are either hydrophobic (H) or polar (P).[140, 141] The HP-sequence is simulated, for instance with a Monte Carlo algorithm, so that positions can be moved on a lattice to adopt a new final conformation, and then contact energy is assigned to all pairs of interactions, depending on the type of interaction found (i.e. HP, PP or HH.)[142, 143, 144, 145] Mutations can also be made to the beads, switching hydrophobic ones into polar or vice-versa. These lattice models have been helpful in illustrating how common neutral changes, defined by energetic thresholds, are distributed within the given model, but these representations are an obvious over simplification of how proteins behave in biology. Amino acids follow more complicated rules, and there are more computing resources available now than when the first lattice models were used to understand proteins.

A more direct comparison between neutral evolution models and biology would need to evaluate tertiary structure, and use realistic amino-acid models while doing so. The DEE/A* protocol would be an ideal solution for this problem: mutations could be evaluated according to how fitness changes as mutations accumulate, as seen in previous chapters. However, rather than compare all mutant sequences to wild type, a different frame of reference could be used instead. Mutations should be compared to the set of neutral sequences that were found, for example, or it may also be reasonable to choose the worst-performing sequence (highest energy, according to the A* search) as a reference point to demonstrate how fitness can be improved from it, continue to a plateau (become neutral) and possibly improve further. A concern here may be in simulating enough sequences to have adequate

sampling for reaction coordinates on the mutagenesis landscape, but even if a single backbone conformation (rather an ensemble of them) is used with DEE/A*, a better resolution for how amino acids behave and interact can be produced.

Adaptive pathways in directed evolution. A similar exploitation of large-scale computational mutagenesis would be in understanding how adaptive evolutionary pathways can be found. In directed evolution experiments, background mutations have an important supporting role in helping target mutations that are expected to improve fitness become fixed into a sequence. These types of experiments may find a few or several combinations of mutations that can achieve a desired level of fitness; DEE/A* can be applied here based on the pathways already found to see if there can be any more. Known mutation pathways should be used as a starting point to calibrate the algorithms as a way to ensure that improvements in fitness can be detected, since biological experiments are available for comparison. Alternative mutations could be limited to positions that are identical to the ones found on these directed-evolution pathways, ones that are in the vicinity and also ones that may be more distant and allosterically linked. Limiting the number of positions would reduce the size of conformational space that needs to be searched. This analysis could potentially provide insight on how much the protein has already evolved. Furthermore, depending on how extensively protein sequence space is explored, situations in which having background mutations as a necessary intermediate step for fixing a desired mutation into a sequence could become more apparent, and DEE/A* could be used then to develop a predictive framework for designing directed evolution experiments.

Bibliography

[1] Rens-Domiano, S. & Hamm, H. E. (1995). Structural and functional relationships of heterotrimeric G-proteins. *FASEB J.* 9, 1059–1066.

[2] Neer, E. J. (1995). Heterotrimeric G proteins: organizers of transmembrane signals. *Cell* 80, 249–257.

[3] Neves, S. R., Ram, P. T., & Iyengar, R. (2002). G protein pathways. *Science*. 31, 1636–1639.

[4] Hampoelz, B. & Knoblich, J. A. (2004). Heterotrimeric G proteins: new tricks for an old dog. *Cell* 119, 453–456.

[5] Farfel, Z., Iiri, T., Shapira, H., Roitman, A., Mouallem, M., & Bourne, H. R. (1996). Pseudohypoparathyroidism a novel mutation in the $\beta\gamma$-contact region of Gsα impairs receptor stimulation. *J. Biol. Chem.* 271, 19653–19655.

[6] Schnabel, P. & Böhm, M. (1996). Heterotrimeric G proteins in heart disease. *Cell. Signal.* 78, 187–198.

[7] Cabrera-Vera Vanhuawe, J., Thomas, T. O., Medkova, M., Preiniger, A., Mazzoni, M. R. and Hamm, H. E., T. M. (2003). Insights into G protein structure, function and regulation. *Endocr. Rev.* 24, 765–781.

[8] Marrari, Y., Crouthamel, M., Irannejad, R., & Wedegaertner, P. B. (2007). Assembly and trafficking of heterotrimeric G proteins. *Biochem.* 46, 7665–7677.

[9] Vanderbeld and Kelly, G. M., B. & Vanderbeld and Kelly, G. M., B. (2000). New thoughts on the role of the $\beta\gamma$ subunit in G protein signal transduction. *Biochem. Cell Biol.* 78, 537–550.

[10] Sprang, S. R. (1997). G protein mechanisms: insights from structural analysis. *Annu. review biochemistry* 66, 639–678.

[11] Wall, M. A., Posner, B. A., & Sprang, S. R. (1998). Structural basis of activity and subunit recognition in G protein heterotrimers. *Structure.* 6, 1169–1183.

[12] Denker, B. M., Neer, E. J., & Schmidt, C. J. (1992). Mutagenesis of the amino terminus of the α subunit of the G protein Go. *J. Biol. Chem.* 267, 6272–6277.

[13] Clapham and Neer, E. J., D. E. (1997). G protein $\beta\gamma$ subunits. *Annu. Rev. Phamacology Toxicol.* 37, 167–203.

[14] Milligan and Kostenis, E., G. & Milligan and Kostenis, E., G. (2006). Heterotrimeric G-proteins: a short history. *Br. J. Pharmacol.* 147, S46—-S55.

[15] Garritsen, A., van Galen, P. J. M., & Simonds, W. F. (1993). The N-terminal coiled-coil domain of β is essential for γ association: a model for G-Protein $\beta\gamma$ subunit Interaction. *Proc. Natl. Acad. Sci.* 90, 7706–7710.

[16] Rebois, R. V., Warner, D. R., & Basi, N. S. (1997). Does subunit dissociation necessarily accompany the activation of all heterotrimeric G proteins? *Cell. Signal.* 9, 141–151.

[17] Klein, S., Reuveni, H., & Levitzki, A. (2000). Signal transduction by a nondissociable heterotrimeric yeast G protein. *Proc. Natl. Acad. Sci.* 97, 3219–3223.

[18] Bünemann, M., Frank, M., & Lohse, M. J. (2003). Gi protein activation in intact cells involves subunit rearrangement rather than dissociation. *Proc. Natl. Acad. Science.* 100, 16077–16082.

[19] Yan, K., Kalyanaraman, V., & Gautam, N. (1996). Differential ability to form the G protein $\beta\gamma$ complex among members of the β and γ subunit families. *J. Biol. Chem.* 271, 7141–7146.

[20] Strathmann and Simon, M. I., M. & Strathmann and Simon, M. I., M. (1990). G protein diversity: a distinct class of α subunits is present in vertebrates and invertebrates. *Proc. Natl. Acad. Science.* 87, 9113–9117.

[21] Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E., & Sigler, P. B. (1996). A 2.0 Åcrystal structure of a heterotrimeric G protein. *Nature.* 379, 311–319.

[22] Stirnimann, C. U., Petsalaki, E., Russell, R. B., & Müller, C. W. (2010). WD40 proteins propel cellular networks. *Trends biochemical sciences* 35, 565–574.

[23] Wettschureck, N. & Offermanns, S. (2005). Mammalian G proteins and their cell type specific functions. *Physiol. Rev.* 85, 1159–1204.

[24] Takesono, A., Cismowsky, M. J., Ribas, C., Bernard, M., Chung, P., Hazard, S. I., Duzic, E., & Lanier, S. M. (1999). Receptor-independent activators of heterotrimeric G-protein signaling pathways. *J. Biol. Chem.* 274, 33202–33205.

[25] Manning, D. R. (2003). Evidence mounts for receptor-independent activation of heterotrimeric G proteins normally in vivo: positioning of the mitotic spindle in C. elegans. *Science. Signal.* 2003, pe35.

[26] Anand-srivastava, M. B. (1997). G-protein and membrane signaling in cardiovascular disease. *Hear. Fail. Rev.* 2, 85–94.

[27] Iwase, M., Uechi, M., Vatner, D. E., Asai, K., Shannon, R. P., Kudej, R. K., Wagner, T. E., Wight, D. C., Patrick, T. A., Ishikawa, Y., Homcy, C. J., & Vatner, S. F. (1997). Cardiomyopathy induced by cardiac Gsα overexpression. *Am. J. Physiol.* 272, H585–H589.

[28] Farfel, Z., Bourne, H. R., & Iiri, T. (1999). The expanding spectrum of G protein diseases. *New Engl. J. Medicine* 340, 1012–1020.

[29] Spiegel, A. M. & Weinstein, L. S. (2004). Inherited diseases involving G proteins and G protein-coupled receptors. *Annu. review medicine* 55, 27–39.

[30] Patten, J. L., Johns, D. R., Valle, D., Eil, C., Gruppuso, P. A., Steele, G., Smallwood, P. M., & Levine, M. A. (1990). Mutation in the gene encoding the stimulatory G protein of adenylate cyclase in Albright's hereditary osteodystrophy. *New Engl. J. Medicine* 322, 1412–1419.

[31] Spiegel, A. M. (1995). Defects in G protein-coupled signal transduction in human disease. *Annu. Rev. Physiol.* 58, 143–170.

[32] Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Science.* 89, 10915–10919.

[33] Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. Chapter 22: A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*, pages 345–352. National Biomedical Research Foundation, Washington D. C., 1978.

[34] Gonnet, G. H., Cohen, M. A., & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science.* 256, 1443–1445.

[35] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

[36] Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217.

[37] Thompson, J. D., Gibson, T. J., Plewniak, F., & Jeanmougin, F. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882.

[38] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

[39] Lapedes, A. S., Giraud, B. G., Liu, L., & Stormo, G. D. (1999). Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Lect. Notes-Monograph Ser.* 33, 236–256.

[40] Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* 286, 295–299.

[41] Süel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature. Struct. Biol.* 10, 59–69.

[42] Magliery, T. J. & Regan, L. (2005). Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics.* 6.

[43] McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., & Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature.* 491, 138–142.

[44] de Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature. Rev. Genet.* 14, 249–261.

[45] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., & Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci.* 106, 67–72.

[46] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* 108, E1293–E1301.

[47] Marks, D. S., Hopf, T. A., & Sander, C. (2012). Protein structure prediction from sequence variation. *Nature. biotechnology* 30, 1072–1080.

[48] Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature. Rev. Genet.* 5, 366–375.

[49] Dean, A. M. & Thornton, J. W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. *Nature. Rev. Genet.* 8, 675–688.

[50] Ortlund Bridgham, J. T., Redinbo, M. R., Thornton, J. W., E. A. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science.* 317, 1544–1548.

[51] Sidhu, S. S. & Koide, S. (2007). Phage display for engineering and analyzing protein interaction interfaces. *Curr. Opin. Struct. Biol.* 17, 481–487.

[52] Ernst, A., Gfeller, D., Kan, Z., Seshagiri, S., Kim, P. M., Bader, G. D., & Sidhu, S. S. (2010). Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. BioSystems* 6, 1782–1790.

[53] Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., & Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Science.* 109, 16858–16863.

[54] Axe, D. D., Foster, N. W., & Fersht, A. R. (1998). A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochem.* 37, 7157–7166.

[55] Cunningham, B. C. & Wells, J. A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science.* 244, 1081–1085.

[56] Weiss, G. A., Watanabe, C. K., Zhong, A., Goddard, A., & Sidhu, S. S. (2000). Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci.* 97, 8950–8954.

[57] Massova, I. & Kollman, P. A. (1999). Computational alanine scanning to probe proteinprotein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* 121, 8133–8143.

[58] Kortemme, T., Kim, D. E., & Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. *Science. Signal.* 2004, pl2.

[59] Street and Mayo, S. L., A. G. & Street and Mayo, S. L., A. G. (1999). Computational protein design. *Structure.* 7, 105–109.

[60] Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science. (New York, N.Y.)* 310, 638–42.

[61] Lippow, S. M. & TIdor, B. (2007). Progress in computational protein design. *Curr. Opin. Biotechnol.* 18, 305–311.

[62] Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science.* 278, 82–87.

[63] Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* 332, 449–460.

[64] Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature.* 453, 190–195.

[65] Desjarlais, J. R. & Clarke, N. D. (1998). Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* 8, 471–475.

[66] Jones, D. T. (1994). De novo protein design using pairwise potentials and a genetic algorithm. *Prot. Science*. 3, 567–574.

[67] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, A. H. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.

[68] Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* 97, 10383–10388.

[69] Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* 383, 66–93.

[70] Li, Z. & Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci.* 84, 6611–6615.

[71] Yang, X. & Saven, J. G. (2005). Computational methods for protein design and protein sequence variability: biased Monte Carlo and replica exchange. *Chem. Phys. Lett.* 401, 205–210.

[72] Hu, X., Hu, H., Beratan, D. N., & Yang, W. (2010). A gradient-directed Monte Carlo approach for protein design. *J. Comput. Chem.* 31, 2164–2168.

[73] Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr. Opin. Biotechnol.* 6, 222–226.

[74] Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* 236, 918–939.

[75] Desmet, J., De Maeyer, M., Hazes, B., & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*. 356, 539–542.

[76] Leach, A. R. & Lemon, A. P. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins: Struct. Funct. Genet.* 33, 227–239.

[77] Gordon, D. B. & Mayo, S. L. (1998). Radical Performance Enhancements for Combinatorial Optimization Algorithms Based on the Dead-End Elimination Theorem. *J. Comput. Chem.* 19, 1505–1514.

[78] Pierce, N. A., Spriet, J. A., & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. *J. Comput. Chem.* 21, 999–1009.

[79] Voigt, C. A., Gordon, D. B., & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* 299, 789–803.

[80] Romero, P. A. & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature. Rev. Mol. Cell Biol.* 10, 866–876.

[81] Bloom, J. D. & Arnold, F. H. (2009). In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci.* 106, 9995–10000.

[82] DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F., & Lombardi, A. (1999). De novo design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* 68, 779–819.

[83] Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L., & Springer, T. A. (2000). Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature. Struct. Biol.* 7, 674–678.

[84] Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* 98, 14274–14279.

[85] Sarkar, C. A., Lowenhaupt, K., Horan, T., Boone, T. C., Tidor, B., & Lauffenburger, D. A. (2002). Rational cytokine design for increased lifetime and enhanced potency using pH-activated "histidine switching". *Nature. Biotechnol.* 20, 908–913.

[86] Looger, L. L., Dwyer, M. A., Smith, J. J., & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*. 423, 185–190.

[87] Bolon, D. N., Grant, R. A., Baker, T. A., & Sauer, R. T. (2005). Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci*. 102, 12724–12729.

[88] Chothia, C. & Lesk, A. M. (1986). The relation bewteen the divegence of sequence and structure in proteins. *The EMBO J*. 5, 823–826.

[89] Chothia, C., Hubard, T., Brenner, S., Barns, H., & Murzin, A. (1997). Protein folds in the all-β and all-α classes. *Annu. review biophysics biomolecular structure* 26, 597–627.

[90] Cordes, M. H. J., Burton, R. E., Walsh, N. P., Mcknight, C. J., & Sauer, R. T. (2000). An evolutionary bridge to a new protein fold. *Nature. Struct. Biol*. 7, 10–13.

[91] Newlove, T., Konieczka, J. H., & Cordes, M. H. J. (2004). Secondary structure switching in Cro protein evolution. *Structure*. 12, 569–581.

[92] Van Dorn, L. O., Newlove, T., Chang, S., Ingram, W. M., & Cordes, M. H. J. (2006). Relationship between sequence determinants of stability for two natural homologous proteins with different folds. *Biochem*. 45, 10542–10553.

[93] Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*. 282, 1462–1467.

[94] Georgiev, I. & Donald, B. R. (2007). Dead-end elimination with backbone flexibility. *Bioinformatics*. 23, i185–i194.

[95] Smith, C. A. & Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol*. 380, 742–756.

[96] Havranek, J. J. & Baker, D. (2009). Motif-directed flexible backbone design of functional interactions. *Prot. science : a publication Prot. Soc*. 18, 1293–1305.

[97] Smith, C. A. & Kortemme, T. (2011). Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PloS ONE* 6, e20451.

[98] Hallen, M. A., Keedy, D. A., & Donald, B. R. (2013). Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. *Proteins* 81, 18–39.

[99] Wall, M. A., Coleman, D. E., Lee, E., Iñiguez Lluhi, J. A., Posner, B. A., Gilman, A. G., & Sprang, S. R. (1995). The structure of the G protein heterotrimer Giα1β1γ2. *Cell* 83, 1047–1058.

[100] Carrascal, N. *Structural and Energetic Determinants of Function in the Heterotrimeric G-proteins*. PhD thesis, Stony Brook University, December 2011.

[101] Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: application to side-chain prediction. *J. Mol. Biol.* 230, 543–574.

[102] Mendes, J., Baptista, A. M., Carrondo, M. A., & Soares, C. M. (1999). Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins: Struct. Funct. Genet.* 37, 530–543.

[103] Green, D. F. (2010). A statistical framework for hierarchical methods in molecular simulation and design. *J. Chem. Theory Comput.* 6, 1682–1697.

[104] Im, W., Lee, M. S., & Brooks, C. L. (2003). Generalized born model with a simple smoothing function. *J. Comput. Chem.* 24, 1691–1702.

[105] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature.* 217, 624–626.

[106] King, J. L. & Jukes, T. H. (1969). Non-Darwiniam evolution. *Science.* 164, 788–798.

[107] Bohm, A., Gaudet, R., & Sigler, P. B. (1997). Structural aspects of heterotrimeric G-protein signaling. *Curr. Opin. Biotechnol.* 8, 480–48.

[108] Conklin, B. R. & Bourne, H. R. (1993). Structural elements of Gα subunits that interact with G$\beta\gamma$, receptors, and effectors. *Cell* 73, 631–641.

[109] Wu, X.-H., Wang, Y., Zhuo, Z., Jiang, F., & Wu, Y.-D. (2012). Identifying the hotspots on the top faces of WD40-repeat proteins from their primary sequences by β-bulges and DHSW tetrads. *PloS One* 7, e43005.

[110] Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Prot. Science*. 3, 211–226.

[111] Archontis, G., Simonson, T., & Karplus, M. (2001). Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol*. 306, 307–327.

[112] Elcock, A. H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol*. 312, 885–896.

[113] Hendsch, Z. S., Nohaile, M. J., Sauer, R. T., & Tidor, B. (2001). Preferential heterodimer formation via undercompensated electrostatic interactions. *J. Am. Chem. Soc*. 123, 1264–1265.

[114] Green, D. F. & Tidor, B. (2005). Design of improved protein inhibitors of HIV-1 cell entry: Optimization of electrostatic interactions at the binding interface. *Proteins: Struct. Funct. Bioinformatics*. 60, 644–657.

[115] Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol*. 8, 338–345.

[116] Pellegrini, M., Marcotte, E. M., & Yeates, T. O. (1999). A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* 35, 440–446.

[117] Marcotte, E. M., Pellegrini, M., Yeates, T. O., & Eisenberg, D. (1999). A census of protein repeats. *J. molecular biology* 293, 151–60.

[118] Andrade, M. A., Perez-Iratxeta, C., & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* 134, 117–131.

[119] Neer, E. J. & Smith, T. F. (1996). G protein heterodimers: new structures propel new questions. *Cell* 84, 175–178.

[120] Smith, T. F., Gaitatzes, C., Saxena, K., & Neer, E. J. (1999). The WD repeat: a common architecture for diverse functions. *Trends biochemical sciences* 24, 181–185.

[121] Wittinghofer, A. (1996). Deciphering the alphabet of G proteins: the structure of the α, β, γ heterotrimer. *Structure.* 4, 357–361.

[122] Fülöp, V. & Jones, D. T. (1999). Beta propellers: structural rigidity and functional diversity. *Curr. opinion structural biology* 9, 715–21.

[123] Neer, E. J., Schmidt, C. J., Nambudripad, R., & Smith, T. F. (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature.* 371, 297–300.

[124] Paoli, M. (2001). Protein folds propelled by diversity. *Prog. biophysics molecular biology* 76, 103–30.

[125] Murzin, A. G. (1992). Structural principles for the propeller assembly of β-sheets: the preference for seven-fold symmetry. *Proteins: Struct. Funct. Genet.* 14, 191–201.

[126] Smith, T. F. (2008). Diversity of WD-repeat proteins. *Sub-cellular Biochem.* 48, 20–30.

[127] Tarnowski, K., Fituch, K., Szczepanowski, R. H., Dadlez, M., & Kaus-Drobek, M. (2014). Patterns of structural dynamics in RACK1 protein retained throughout evolution: a hydrogen-deuterium exchange study of three orthologs. *Prot. Science.* 23, 639–651.

[128] Springer, T. A. (1997). Folding of the N-terminal, ligand-binding region of integrin α-subunits into a β-propeller domain. *Proc. Natl. Acad. Sci.* 94, 65–72.

[129] Wu, X.-H., Chen, R.-C., Gao, Y., & Wu, Y.-D. (2010). The effect of Asp-His-Ser/Thr-Trp tetrad on the thermostability of WD40-repeat proteins. *Biochem.* 49, 10237–10245.

[130] Wu, X.-H., Zhang, H., & Wu, Y.-D. (2009). Is Asp-His-Ser/Thr-Trp tetrad hydrogen-bond network important to WD40-repeat proteins: a statistical and theoretical study. *Proteins* 78, 1186–1194.

[131] Wang, Y., Jiang, F., Zhuo, Z., Wu, X.-H., & Wu, Y.-D. (2013). A method for WD40 repeat detection and secondary structure prediction. *PloS One* 8, e65705.

[132] Kloss, E., Courtemanche, N., & Barrick, D. (2008). Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. biochemistry biophysics* 469, 83–99.

[133] Barrick, D., Ferreiro, D. U., & Komives, E. A. (2008). Folding landscapes of ankyrin repeat proteins: experiments meet theory. *Curr. opinion structural biology* 18, 27–34.

[134] Sikorski, R. S., Boguski, M. S., Goebl, M., & Hieter, P. (1990). A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell* 60, 307–317.

[135] Sedgwick, S. G. & Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends biochemical sciences* 24, 311–6.

[136] Lux, S. E., John, K. M., & Bennett, V. (1990). Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. *Nature.* 344, 36–42.

[137] Li, D. & Roberts, R. (2001). WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell. Mol. Life Sci.* 58, 2085–2097.

[138] Chen, C. K.-M., Chan, N.-L., & Wang, A. H.-J. (2011). The many blades of the β-propeller proteins: conserved but versatile. *Trends Biochem. Sci.* 36, 553–561.

[139] Ashkenazy, H., Unger, R., & Kliger, Y. (2009). Optimal data collection for correlated mutation analysis. *Proteins: Struct. Funct. Bioinformatics.* 74, 545–555.

[140] Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochem.* 24, 1501–1509.

[141] Lau, K. F. & Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromol.* 22, 3986–3997.

[142] Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophys. J.* 73, 2393–2403.

[143] Bastolla, U., Roman, H. E., & Vendruscolo, M. (1999). Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* 200.

[144] Bornberg-Bauer and Chan, H. S., E. & Bornberg-Bauer and Chan, H. S., E. (1999). Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Science.* 96, 10689–10694.

[145] Bastolla Porto, M., Roman, H. E., and Vendruscolo, M., U. & Bastolla Porto, M., Roman, H. E., and Vendruscolo, M., U. (2003). Connectivity of neutral networks, overdispersion, and structural conservation in protein evolutions. *J. Mol. Evol.* 56, 243–254.