Computational Design of Protein Structure and Prediction of Ligand Binding

by

Robert Aron Broom

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Chemistry

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Proteins perform a tremendous array of finely-tuned functions which are not only critical in living organisms, but can be used for industrial and medical purposes. The ability to rationally design these molecular machines could provide a wealth of opportunities, for example to improve human health and to expand the range and reduce cost of many industrial chemical processes. The modularity of a protein sequence combined with many degrees of structural freedom yield a problem that can frequently be best tackled using computational methods. These computational methods, which include the use of: bioinformatics analysis, molecular dynamics, empirical forcefields, statistical potentials, and machine learning approaches, amongst others, are collectively known as Computational Protein Design (CPD). Here CPD is examined from the perspective of four different goals: successful design of an intended structure, the prediction of folding and unfolding kinetics from structure (kinetic stability in particular), engineering of improved stability, and prediction of binding sites and energetics.

A considerable proportion of protein folds, and the majority of the most common folds ("superfolds"), are internally symmetric, suggesting emergence from an ancient repetition event. CPD — an increasingly popular and successful method for generating *de novo* folded sequences and topologies — suffers from exponential scaling of complexity with protein size. Thus, the overwhelming majority of successful designs are of relatively small proteins (< 100 amino acids). Designing proteins comprised of repeated modular elements allows the design space to be partitioned into more manageable portions. Here, a bioinformatics analysis of a "superfold", the β-trefoil, demonstrated that formation of a globular fold *via* repetition was not only an ancient event, but an ongoing means of generating diverse and functional sequences. Modular repetition also promotes rapid evolution for binding multivalent targets in the "evolutionary arms race" between host and pathogen. Finally, modular repetition was used to successfully design, on the first attempt, a well-folded and functional β-trefoil, called ThreeFoil.

Improving protein design requires understanding the outcomes of design and not simply the 3D structure. To this end, I undertook an extensive biophysical characterization of ThreeFoil, with the key finding that its unfolding is extraordinarily slow, with a half-life of almost a decade. This kinetic stability grants ThreeFoil near-immunity to common denaturants as well as high resistance to proteolysis. A large scale analysis of hundreds of proteins, and coarse-grained modelling of ThreeFoil and other beta-trefoils, indicates that high kinetic stability results from a folded structure rich in contacts between residues distant in sequence (long-range contacts). Furthermore, an analysis of unrelated proteins known to have similar protease resistance, demonstrates that the topological complexity resulting from these long-range contacts may be a general mechanism by which proteins remain folded in harsh environments.

Despite the wonderful kinetic stability of ThreeFoil, it has only moderate thermodynamic stability. I sought to improve this in order to provide a stability buffer for future functional engineering and mutagenesis. Numerous computational tools which predict stability change upon point mutation were used, and 10 mutations made based on their recommendations. Despite claims of >80% accuracy for these predictions, only 2 of the 10 mutations were stabilizing. An in-depth analysis of more than 20 such tools shows that, to a large extent, while they are capable of recognizing highly destabilizing mutations, they are unable to distinguish between moderately destabilizing and stabilizing mutations.

Designing protein structure tests our understanding of the determinants of protein folding, but useful function is often the final goal of protein engineering. I explored protein-ligand binding using molecular dynamics for several protein-ligand systems involving both flexible ligand binding to deep pockets and more rigid ligand binding to shallow grooves. I also used various levels of simulation complexity, from gas-phase, to implicit solvent, to fully explicit solvent, as well as simple equilibrium simulations to interrogate known interactions to more complex energetically biased simulations to explore diverse configurations and gain novel information.

Acknowledgements

Beyond simply “standing on the shoulders of giants”, I have been able to complete the research within this thesis, thanks to the steadying hands placed upon me (giants have rather uneven shoulders after all) by my: colleagues, supervisor, family, and most importantly of all, my wife, Helen.

Though I note here that Poppy, our genetically and intellectually disadvantaged dog/raccoon, has had at least some small part to play in all of this.

Dedication

In the TV series *Star Trek: the Next Generation*, humans travel around on spaceships so massive they accommodate the population of a small city, families, children, pets and all. These ships move at many times the speed of light (at “warp 9”, the ship moved at 9^3 times the speed of light), can be cloaked from detection, repel all manner of projectiles and energy beams, can pull distant objects without making physical contact, and can search an entire planet for life in a matter of seconds. The computer systems on these ships are nearly sentient, and can easily store, access, and manipulate the sum of human (and other intelligent species’) knowledge. The people themselves can be teleported between distant locations in a matter of seconds, their bodies broken down into energy and that energy reassembled into mass with precisely the right atomic level structure hundreds of kilometers away (conveniently this can also be used to make food). And yet, the oldest human being in *Star Trek: the Next Generation*, at 137, was Leonard McCoy, more familiarly known as Dr. McCoy or “Bones”.

How is it that such a deep science fiction universe, while showing us a vision of the future with these extreme advancements in technology, can have so little imagination for the extension of human lifespan? Warp-speed allows travel at more than 10,000 times what is possible today. Information is sent through “sub-space” at many times the speed of light, and therefore many times faster than what is possible now. The iconic “tricorder”, can detect diseases, bone fractures, and material compositions in a way that makes even the most newsworthy medical diagnostic uses of cell-phones look primitive. Food is generated exactly to taste in seconds with no fanfare. But, human lifespan is predicted to be a mere 20 years longer than it is now (the oldest man died 3 years ago at age 116). This disparity originates from problems with simplistic extrapolation and a lack of vision.

Approximately 200,000 years ago humans moved from place to place on foot, a good jog yielding about 10 km/h. About 6000 years ago, we tamed horses, and could move across long distances at speeds of 40 km/h. Around the same time, sailing ships were invented, and while they couldn’t travel faster than a horse, they could move across water, required no rest, and could move heavy loads. 200 years ago, in the age of steam, trains were invented which could reach speeds of 100 km/h, carrying massive loads, but were limited to fixed paths. About 150 years ago the combustion engine allowed motor vehicles to achieve the same speeds with substantially fewer restrictions on path. Then, 100 years ago, we moved to the skies (no path restrictions here), with some of the first airplanes already reaching speeds of 200-300 km/h. The invention of rocket engines 50 years ago allowed travel by air at speeds of more than 1000 km/h. Looking back on the rapid improvement in travel speed, it is easy to understand how one would envision this technological progress to yield faster than light travel within 300 years (*Star Trek* is set in the 24th century). Similar extrapolations can be made for food production and processing, medical diagnostics, and information technology. But what about life-span? While average lifespan has been improving since the days we got on horseback, and is largely attributable to no longer using those same horses to ride up to, and smash in, the head of another human. Maximum lifespan has changed little. 1500 years ago, in ancient Greece, the maximum lifespan of intellectually renowned men — who typically did not have their head's smashed in — was about 80 years. Today, if we similarly examined men of substantial intellectual accomplishment (Nobel prize winners for instance), we see maximum lifespans of at best $\sim$90 years. As mentioned earlier, the oldest man that has ever lived (ignoring Methuselah and other mythological tales), died 3 years ago at the age of 116. Thus, while the ancient Greeks were moving around at 40 km/h on horseback and could live to 80 years of age, we have gained a pathetic 10 or at most 40 years of extra life (50% better at most), while being able to move $\sim$1000 km/h faster (2500% better). As such, it is easy to see why the idea of extremely long lifespans is hardly part of popular science fiction, our own cultural imagination, and consequently, our own cultural desire.

But the invention of antibiotics, while perhaps little changing maximum lifespan (though it did boost the average), showed us that diseases of the extremely complex molecular machines that are our bodies, can be stopped. In fact, determination of the structure and function of DNA, the structure and dynamics of proteins, and their complex interplay with one another and with lipids and carbohydrates, has shown us that these machines can be controlled and potentially repaired, even at the finest level of detail. While progress in the speed at which we travel took many thousands of years, recent similar explosions in capability, such as the many orders of magnitude jump in computational power since the first computers 60 years ago, shows that once we finally start to make progress, it could be extraordinarily rapid. This thesis is dedicated to that goal. That at some time in the near future, our constant struggle to stay alive to experience one more day, will become trivial. And once that happens, then, we can start to trek amongst the stars.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACO - absolute contact order
ANS - anilinonaphthalenesulfonic acid
BSA - bovine serum albumen
BTB - Bis-tris buffer
CC - clustering coefficient
CD - circular dichroism
CDD - conserved domain database
CHCA - α-cyano-4-hydroxcinnamic acid
CPD - computational protein design
CO - contact order
COG - center of geometry
D2D - diagonal two dimensional
DLS - dynamic light scattering
DSC - differential scanning calorimetry
EDTA - ethylenediaminetetraacetic acid
EP - empirical potential
FES - free energy surface
FGF - fibroblast growth factor
FN - false negative
FP - false positive
GPGPU - general purpose graphical processing unit
GS - gold standard
GuHCl - guanidine hydrochloride
GuSCN - guanidine thiocyanate
H/D - hydrogen / deuterium
HMM - hidden Markov model
IPTG - isopropyl β-D-1-thiogalactopyranoside
IQR - inter-quartile range
I/O - input/output
KSP - kinetically stable protein
LC - liquid chromatography
LIE - linear interaction energy

LRO - long range order
MALDI - matrix assisted laser desorption ionization
MCC - Matthews correlation coefficient
MD - molecular dynamics
MLR - Morse/long-range potential
MM - molecular mechanics
MRR - multiple replica repulsions
MS - mass spectrometry
MSA - multiple sequence alignment
NCBI - National Center for Biotechnology Information
NOESY - nuclear Overhauser effect spectroscopy
NMR - nuclear magnetic resonance
PAGE - polyacrylamide gel electrophoresis
PDB - protein data bank
PP - physical potential
PPV - positive predictive value
QM - quantum mechanics
RCO - relative contact order
RCSB - research collaboratory for structural bioinformatics (see also PDB)
RESP - restrained electrostatic potential
RGR - replica ghost repulsions
RMSD - root mean squared deviation
SASA - solvent accessible surface area
SDS - sodium dodecyl sulfate
SEC - size exclusion chromatography
SLS - static light scattering
SP - statistical potential
SSL - secure socket layer
SOD - human Cu,Zn superoxide dismutase
SRO - short range order
TCD - total contact distance
TI - thermodynamic integration
TN - true negative
TOF - time of flight
TP - true positive
TST - transition state theory
VMO-I - vitelline outer membrane protein I
WHAM - weighted histogram analysis method

WT - wild type

Chapter 1

Introduction

1.1 Structure of the thesis

This thesis is constructed from unpublished data and published papers. The published papers are presented in a "pre-print" format, so as to create a cohesive style throughout. At the beginning of each chapter a brief "Context" section attempts to place the chapter in the context of the whole thesis, and for previously published works, delineates my contribution. This introductory chapter is primarily a pre-print of a review "Using natural sequences and modularity to design common and novel protein topologies", recently published in Current Opinion in Structural Biology [1], though it is bracketed by additional short sections in order to introduce the full scope of the thesis.

1.2 Why proteins matter

An American Presidential candidate was once heard to utter, "Proteins do things. Proteins do the best things". While we have moved beyond the early cellular biology dogma suggesting that proteins perform all the enzymatic functions for a cell, it is nevertheless the case that proteins are responsible for the vast majority of biocatalysis [2]. Their capacity for finely tuned specificity in molecular interactions also makes them the dominant players in sensing and recognition and binding events [2]. Thus, many hereditary diseases result from mutations within protein coding regions [3].

The catalytic dominance of proteins has made them key targets in evolving more effective means of biofuel production [4, 5], bioremediation [4, 6], and the field of green chemical synthesis both for common and novelty chemical compounds [4, 7]. The capacity for tight and specific binding has been exploited in the development of many antibodies or antibody mimetics [8]. Such binding abilities have also been used to probe internal cellular environments [9], or to screen and characterize cellular products such as cancer markers [10]. Thus, while proteins already provide valuable biomedical and biocatalytic functions, we have typically only co-opted existing functions, and considerably more may be available form these molecular machines if we could design them with precision and ease.

The finely tuned function of proteins owes itself to their modular form which allows nearly infinite possibilities. Specifically, proteins are linear combinations of amino acids linked by peptide bonds. Depending on the choice of amino acids in this "primary structure", the chain may adopt particular local configurations, known as "secondary structure", and subsequently could adopt a specific final shape or "tertiary structure". Typically it is this shape that provides a protein with its function. In most organisms (ourselves included) there are 20 possible amino acids that can be used, and most protein domains are in the range of ~100-200 amino acids [11]. Thus, even for small proteins there are 20^{100} possible combinations, or a number so large that if we made only a single protein molecule for each sequence, the resulting protein slurry (even without being solvated in water) would be so massive that, placed next to the known universe, it would appear as the Sun placed next to a grain of sand. While this is absolutely marvellous for those who love proteins, it has the unfortunate effect of making the design of these molecules quite complicated.

The complication is increased by the fact that even very similar protein sequences can have different functions [12], and some sequences which are nearly identical nevertheless adopt different tertiary structures or folds [13]. Therefore, a very fine-grained understanding of protein structure and function is needed for effective protein design. Approaches which have been used to tackle the design of a protein's fold are introduced in Section 1.3 with limited emphasis on function. A focus on the use of advanced computational methods such as molecular dynamics to address design of binding function in particular is then introduced in Section 1.4.

1.3 Computational protein design

Protein design has advanced tremendously over the last several decades; yet, the reliable design of a stable, well-folded, and soluble protein with the intended structure remains far from routine, often requiring multiple attempts, iterative improvements, and substantial resources [14, 15]. On the other hand, nature has successfully explored a great diversity of sequences and topologies [12], offering large and rapidly growing repositories of information that are increasingly leveraged in design. The advent of computational protein design (CPD) enabled the exploration of fully *de novo* sequences for natural topologies [16, 17], and more recently the design of *de novo* topologies (also using natural structural information) [18, 19]. While nature's existing sequences and topologies offer a solid foundation for protein design, recent breakthroughs generating natural and novel topologies, often with very high stability, demonstrate that many forms are possible [18, 20, 21, 22, 23, 24].

Much progress has been made concurrently using natural mechanisms to improve or guide design by selection and directed evolution approaches [25]. Here we focus predominantly on rational design of protein topology, highlighting recent developments that leverage sequence and topology databases in addition to CPD using both atomistic energy functions and coarse-grained simulations. Recent results demonstrate the numerous and increasingly sophisticated strategies, combining multiple approaches, which are being developed and validated; these promise to further advance fundamental understanding of the interplay between sequence, folding and topology and to improve the success of protein design in practical applications (Figure 1.1). The ongoing improvements may be likened to the development of refined and powerful machine tools at the onset of the industrial revolution, and may usher in a similarly transformative period.

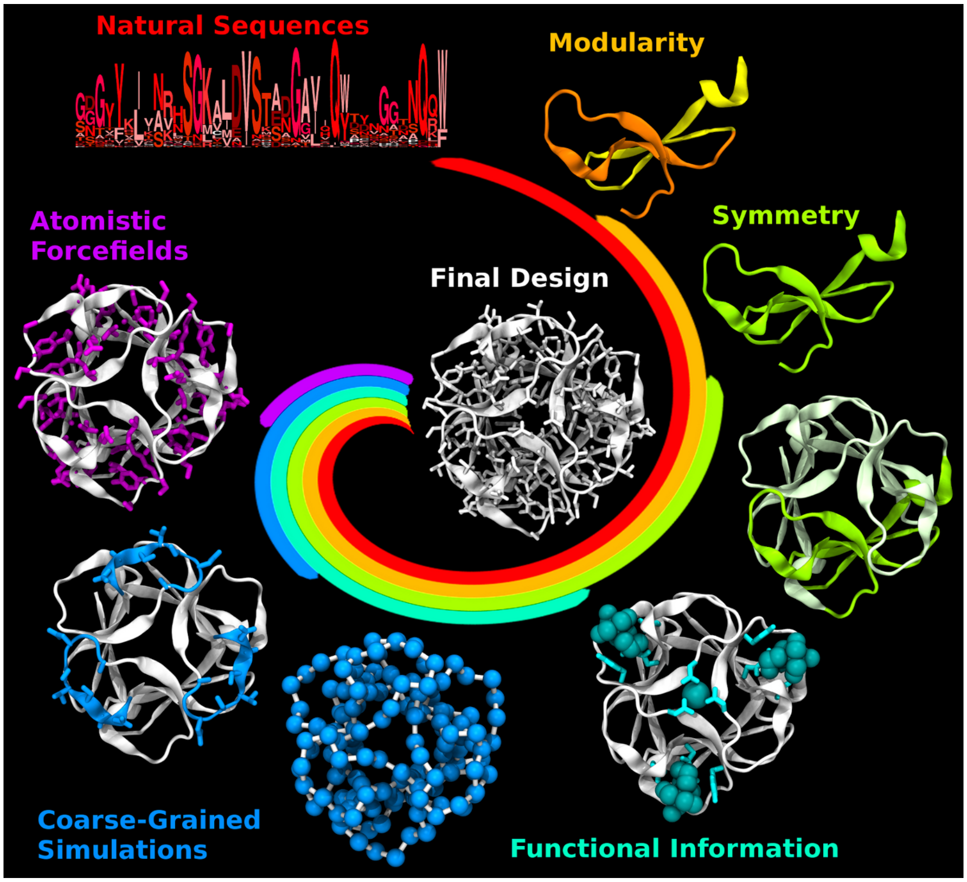


Figure 1.1: Overview of approaches for protein design. The central colored spirals depict the contributions of different design methods, which may vary from one approach to another, and together generate the final design. For each method, the corresponding coloured elements of structure are designed, and may then be fixed in later design stages (white). A wide range of protein topologies have been realized, using approaches based on: information derived from natural sequences, such as consensus (red); small subdomain-sized structural modules (yellow, orange); repetition of structural modules, i.e. symmetry (green); consideration of functional sites (cyan, ligands shown as spheres); coarse-grained simulations (blue spheres represent individual residues in simulations which inform the design of certain portions of the backbone or sidechains); and atomistic simulations using force fields (purple, modelled sidechains shown as sticks). Recent successful designs typically incorporate multiple methods. For instance, incorporating structural modularity (yellow, orange) with symmetry (green) can greatly reduce the size and complexity of the design problem. Including functional constraints (cyan) and/or coarse-grained folding simulations of many or selected residues (blue) can help retain function or optimize folding. The design methods are illustrated using ThreeFoil (PDB: 3PG0).

1.3.1 Natural sequence statistics in design

Sequence data, without explicit consideration of structure, are now widely employed in designing stable and functional proteins. Numerous studies have shown that consensus sequence-based design — choosing the most common amino acid at each position of a multiple sequence alignment (MSA) — can be useful for increasing protein stability and may also aid in diversifying function. Recently, for example, the consensus design of FN3 domains using only a handful of closely-related sequences resulted in significantly increased stability [26], while using thousands of more distantly-related sequences produced an extraordinarily stable variant [27]. In the case of proteins with a catalytic function, large and diverse MSAs similarly resulted in improved stability but often with concurrent loss of catalytic activity and increased substrate promiscuity [28, 29]. A similar loss of specificity may occur when consensus designing proteins with ligand-binding, transport, and other functions. Reconstructing ancestral sequences from the same MSAs used for consensus design, Risso et al. found a similar trend in activity, though the ancestral reconstructions in this case were all more stable than the consensus designs [29]. A cogent review of ancestral sequence reconstruction is provided by Wheeler *et al.* [30]. Sequence-based protein design may be more successful for highly populated folds which provide larger alignments, and such folds may be highly populated precisely because they are more amenable to functional diversification or engineering [31, 32].

The success of consensus design is proposed to arise from the tendency for natural sequences to drift over time while the amino acids critical for stability and folding are retained owing to evolutionary pressure [33]. This phenomenon can be utilized to improve the folding and stability of natural and designed proteins by making specific "consensus mutations" [34]. Further, the covariation or correlations between amino acids at different positions in an MSA may be useful for structure prediction and identifying functional residues [35, 36]. Accounting for covariation when choosing consensus mutations may improve stability while also reducing potentially negative impacts on function [34, 37]. Entirely sequence-based design methods have been successfully applied and continue to be developed; in more complex designs, inclusion of sequence-based tools has become common and promises to be ever more widely useful (Figure 1.1).

1.3.2 Modular topology simplifies design

Sequence-based methods are often included as part of more complex design approaches that have been applied to make a wide range of topologies. For instance, various *de novo* sequence designs have been successful for small folds (∼100 amino acids or smaller)

[16, 17, 18, 38, 39]. The design of larger and often more complex topologies has been aided by using the modular nature of protein structures [40, 41] to simplify the design process and improve the tractability of CPD [42]. In particular, various common protein topologies have evolved *via* duplication(s) of a single module [40, 43] (Figure 1.2), and design of a single subdomain module (of ∼20-40 amino acids) repeated to form a larger domain, is an attractive means of simplifying the design process [21, 22, 23, 44, 45, 46, 47]. Here we use the term repetition from a structural perspective, referring to domains which can be deconstructed into repeated subdomain modules, making no assumptions concerning the genetic mechanisms giving rise to such domains. Furthermore, the term subdomain module refers to a peptide with continuous sequence that forms a compact structural element that is smaller than an autonomously folding domain. Notably, ∼20% of all folds are estimated to have internal repetition or symmetric structures, including a substantial proportion of the topologies most frequently observed for natural proteins [43, 48]. We begin by examining the recombination of subdomain modules and follow with special cases where a single module is repeated to generate a larger whole domain. Both types of fold/domain generation have been successful as rational design approaches, but have also been used by nature to successfully generate novelty and diversity. Thus, natural protein evolution can inform how one approaches design [31, 49, 50] and conversely, design studies can provide insight into evolution [44, 45, 51, 52, 53, 54].

Recombination between modules

Topological or structural modularity has been leveraged by recombining sequences corresponding to the same module from different proteins. Chimeras made of modules chosen so as to disrupt a minimal number of residue-residue contacts [55] are quite frequently folded and functional [56]. Chimeric proteins can combine desirable properties of both parents; for instance, recombination of a protein with high stability and another with a desired function (but less stability) can frequently result in improved stability while retaining function [57, 58, 59]. Large changes in sequence (particularly at modular interfaces) accompanying recombination may also alter structural dynamics. Notably, Gobeil *et al.* found for TEM-1 β-lactamase that while the two parent enzymes had similar dynamics, those of the chimera were dramatically increased; yet, substrate turnover was unaltered [60]. Thus, while dynamics may be critical for function, altered dynamics need not limit design and may even foster innovating useful new functional features, particularly in the context of a robust scaffold [32].

Beyond substituting homologous sequences, combining modules with unrelated sequences may offer opportunities for even greater diversification and innovability. Experimental and theoretical analyses have demonstrated that non-homologous recombinations may tunnel under barriers insurmountable *via* point mutations to create novel proteins [61, 62]. To date, examples of rational design using non-homologous recombination have been rare, perhaps because disruption of correlated amino acids at module interfaces reduces the likelihood of success. In one case, Bharat *et al.* attempted to build a new $(\beta\alpha)_8$-barrel using half of a known $(\beta\alpha)_8$-barrel and a portion of a very distantly related protein with a $(\beta\alpha)_5$-flavodoxin-like topology. The chimera was stable, monomeric, and exhibited cooperative unfolding, but adopted an unintended 9-stranded barrel structure [51]. Subsequently, Rosetta was used to make five interface optimizing mutations, resulting in the original target topology and increased stability [63]. Thus, the modular nature of protein topologies may help promote generating folded and functionally diversified proteins, but more adventurous recombinations between distantly or unrelated modules will likely require complementary strategies — such as CPD — to maximize success.

Linear repeat proteins

The repetition of a single subdomain module to construct an entire protein has been studied extensively for linear repeat proteins, which have elongated structures consisting of topologically identical modules (Figure 1.2a,b). The regularity of repeat protein structures offers advantages for designing versatile binding scaffolds [64, 65], and their linear character simplifies understanding and modelling of their folding behaviour [66]. modelling has suggested that with sufficient interaction energies between modules, longer repeat proteins can be extremely stable. This is exemplified by a consensus designed ankyrin repeat protein comprised of three identical internal repeats (plus distinct N- and C- cap repeats) having a T_m >100°C, with additional internal repeats resulting in variants resistant to unfolding by high temperature alone [67]. Coarse-grained simulations have shown that a higher ratio of inter- vs. intra-repeat contacts results in slower and more cooperative kinetic transitions, explaining the kinetic stability of this design [68]. Consistent with the simulations, additional intra-repeat contacts increased the experimentally determined folding cooperativity for a redesigned repeat protein [69], and may be pertinent to creating high cooperativity and kinetic stability in other protein topologies [70, 71].

Combining the Rosetta atomistic forcefield with natural sequence statistics, Parmeggiani *et al.* redesigned five repeat protein folds as well as three types of β-propeller structures (see "Toroidal β-propellers"). Excepting the β-propellers, the designs were largely successful with 60% forming folded monomers, many with a T_m above 95°C [47]. In a striking application of rational design, Park *et al.* used Rosetta to make a series of repeat modules that could be intermixed to fine-tune the curvature of the resulting structures, providing rational control over shape complementarity for binding [22]. In addition to optimization of binding interfaces through shape complementarity, some binding targets may require unique modules with highly correlated sequence choices providing a coordinated binding surface [72]. The above examples illustrate how linear repeat proteins have emerged as particularly designable; lessons learned about the molecular determinants of their kinetic stability and cooperativity may be extendable to other modular protein topologies, considered further below.

Toroidal β-propellers

β-propellers can be thought of as a repeat protein with interacting N- and C-terminal modules forming a toroidal structure (often with a solvated core) (Figure 1.2c). Consequently, toroidal structures may be more challenging design targets than linear-repeat proteins with fewer constraints on interactions between modules. In the work of Parmeggiani *et al.* (see "Linear repeat proteins"), 13 monomeric β-propeller designs of 6-, 7- and 8-bladed β-propellers were tested experimentally. Only a single design appeared folded, but formed an unintended dimer, hence, success was markedly lower for these toroidal structures than for the linear repeats [47]. Nevertheless, *de novo* β-propeller sequences have been successfully designed using entirely natural sequences and in combination with CPD. Smock *et al.* performed ancestral reconstruction starting with a highly symmetric natural 5-blade β-propeller to generate a putative ancient single blade. Based on this module they demonstrated a plausible evolutionary path from a functional homopentamer to a monomeric, symmetric and functional 5-bladed β-propeller [54]. Starting from another highly symmetric natural β-propeller (in this case 6-bladed), Voet *et al.* used limited consensus design combined with Rosetta-based CPD with symmetry constraints to design a single blade module. Proteins containing 2-10 identical repeats of this blade formed 6-bladed structures within various oligomeric assemblies [45]. Subsequently, engineering of a metal binding site into 2-bladed monomers resulted in 6-bladed assemblies stabilized by metals and promoting nanocrystal formation [73]. These studies further illustrate benefits of simplifying the design process by using repeated modules, and demonstrate the utility of combining natural sequence information to test evolutionary hypotheses as well as produce successful designs.

Globular folds with internal symmetry

Compared with linear repeat and toroidal proteins, globular proteins tend to have increased structural complexity (owing to their need for a well-packed hydrophobic core), but many are thought to have similarly arisen from the repetition of smaller subdomain modules [40, 43, 48]. Early tests of this idea focused on the ubiquitous $(\beta\alpha)_8$-barrel "superfold" [48], by designing structures proposed to have evolved from the repetition of half- or quarter-barrels [52] (Figure 1.2d). These investigations culminated in the design of several stable and soluble $(\beta\alpha)_8$-barrels composed of identical half-barrel [74, 75] or disulfide-linked quarter-barrel modules [53]. While these $(\beta\alpha)_8$-barrel cases used nature-sourced sequences as templates, several recent studies have undertaken *de novo* design using Rosetta. Despite numerous computational checks and balances, the *de novo* design of a 216-residue $(\beta\alpha)_8$-barrel, while adopting the intended secondary structure, folded non-cooperatively and was poorly soluble [76]. Taking advantage of symmetry by designing identical quarter-barrel modules using Rosetta, and incorporating rules for idealized backbones [39], Nagarajan *et al.* [77] and Huang *et al.* [46] both generated soluble designs. The design by Nagarajan *et al.* had marginal thermodynamic stability (∼2 kcal/mol) and NMR suggested a molten globule [77]. On the other hand, the design of Huang *et al.* (which imposed additional rational design constraints on the backbone structure, loop residues, and hydrophobicity of the barrel interior) produced the first *de novo* sequence structurally confirmed to adopt the intended $(\beta\alpha)_8$-barrel topology. Still, the thermodynamic stability was moderate (∼4 kcal/mol) [46], considering that larger structures can have increased capacity for high thermodynamic stability [21, 67, 69, 78].

Another internally symmetric "superfold", the β-trefoil, consists of three repeats of a four β-strand module (Figure 1.2e). Interestingly, analysis of β-trefoil sequences indicates that the repetition of modules to generate new folds was not only a common ancient occurrence [40, 43, 52] but can be ongoing [44]. Our group and the Blaber group independently designed fully symmetric β-trefoils, called ThreeFoil and Symfoil, using very different approaches. For ThreeFoil, starting with a highly symmetric predicted β-trefoil sequence as a template and combining consensus design using close homologs (to preserve function) with Rosetta-based CPD, the multivalent carbohydrate binding target was obtained in a single attempt [44]. Multivalent binding is a common feature of proteins with modular repeats as in this case as well as toroidal β-propellers [54] and linear repeat proteins [65, 66, 69, 72]. ThreeFoil has significant thermodynamic stability and remarkably high kinetic stability (unfolding half-life ∼8 yrs) as well as high resistance against proteolytic degradation, chaotropes and detergent [70]. Analysis using contact order and coarse-grained modelling indicates the extremely slow unfolding arises from a highly cooperative topology containing numerous long range contacts [70, 78]. Incorporating desired kinetic stability is a little-explored but promising avenue for future protein designs.

Using an iterative process, Symfoil was designed starting from the natural fibroblast growth factor (FGF), with many rounds of design and selection gradually increasing sequence symmetry [79]. This process eliminated function but resulted in high thermodynamic stability. While the design of Symfoil and ThreeFoil used the structural module inferred from sequence analysis of natural proteins, an alternate FGF-based design used a module comprising the experimentally determined folding nucleus. This design also ablated function, but exhibited improved solubility and stability compared with FGF, suggesting the repeated topological module need not be defined by inferred evolutionary pathways [80].

Thus, the repetition of a simple topological module has proven successful for globular folds despite their increased structural complexity, including long-range contacts between symmetric residues.

Parametric design of helices

The complex nature of protein topologies in addition to the interactions of numerous residues in different structural contexts typically requires in an extremely large and complex design space. By contrast, the α-helical coiled-coil topology is particularly amenable to design using relatively few parameters [81]. Recent advances have leveraged the combined strengths of parametric design and CPD to realize a range of supercoiled α-helical architectures [19, 21].

Classical coiled-coils, long a target of protein design, are assemblies of α-helices containing heptad sequence repeats [82], and oligomers composed of five or more helices are described as α-helical barrels, which may enclose a central channel [19] (Figure 1.2f). Using structure-based computational methods to screen candidate heptad sequences, Thomson *et al.* rationally designed repeats compatible with specific α-helical barrel parameters, including oligomer state, thereby tuning the size of the central channel [19]. The rational design of cage-like structures has also been accomplished using heptad-repeat coiled-coils, both as monomers [83] and oligomers [84]. Using longer repeats of 11- and 18-residues with Rosetta, Huang *et al.* made 4- and 3-helix bundles consisting of 191 and 247 amino acids, respectively (Figure 1.2g); the clever use of repetition within the long helices and symmetry between helices greatly reduced the number of configurations for structure-based evaluation [21]. These relatively large monomeric proteins exhibited extremely high stability (e.g. extrapolated $\Delta G > 61$ kcal/mol).

Taken together, the studies highlighted herein show how comparatively small, repeated modules simplify the design process and have yielded many first round successes for a wide range of topologies [19, 21, 45, 46, 70, 85]. As many of the most common protein folds exhibit internal symmetry [43] there are many opportunities to take advantage of symmetry and repetition in design.

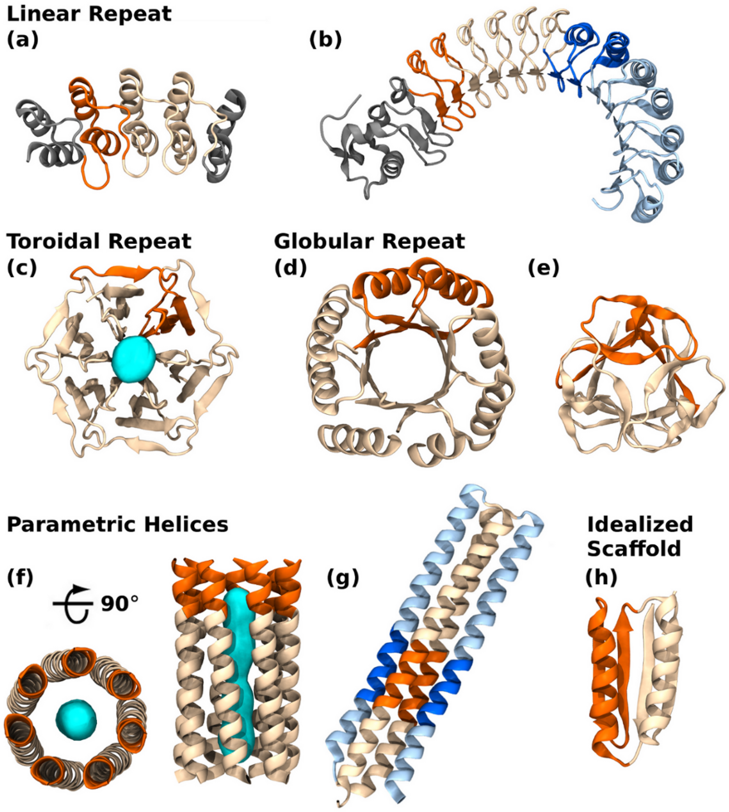


Figure 1.2: Protein topology designs taking advantage of modularity and symmetry. Representative structures of recent successful designs are shown with a single modular element highlighted in dark orange/blue. The remaining part of the structure constructed by repeating the single element is shown as light orange/blue. Linear repeat proteins can be made by repetition of (a) topologically identical modules [47] resulting in uniform curvature, as typically observed in nature, or (b) using different modules (orange and blue) to give rational control over curvature for binding interactions [22]. For the linear repeat proteins, capping modules (grey) for improving protein solubility have a modified sequence and may have (a) the same or (b) different structures as the other repeats. Toroidal proteins such as (c) the β-propeller [45] are distinct from linear repeat proteins in that the N- and C-terminal repeats interact, forming a continuous topology without the need for capping modules, and may enclose a solvated pore (cyan). Globular proteins such as (d) the 4-fold symmetric $(\beta\alpha)_8$-barrel [46] and (e) the 3-fold symmetric β-trefoil [70, 79] have interacting N- and C-terminal repeats like the toroidal proteins, but with a central hydrophobic core. Parametrically designed helices forming coiled-coils may be designed as (f) homo-oligomers or (g) monomers, with each helix generated from the repetition of a small sequence motif. The homo-oligomers designed by Thomson *et al.* have the potentially useful property of containing a central solvated channel (cyan surface) with a rationally designable diameter [19]. In the extremely stable monomeric helical bundle [21] two different motifs (orange and blue) are repeated to form a structure with 2-fold symmetry (g). Several idealized scaffolds [39], while not designed using modular repetition, nevertheless have structures with significant symmetry (h), a property of many common protein folds [43, 48], suggesting such structures are inherently amenable to design. The PDB IDs for each structure are: a) 2XEE, b) 4R5D, c) 3WW9, d) 5BVL, e) 3PG0, f) 4PNA, g) 4UOS, and h) 2KL8.

1.3.3 Improving designs

Natural, well-folded proteins typically have native states that have significant but not very high stability compared to their unfolded states (∼3-10 kcal/mol). Because the stabilization resulting from the burial of a hydrophobic group or the formation of a hydrogen bond can be on the order of ∼1 kcal/mol [86], even small improvements, for example in packing efficiency [87], can substantially improve protein design outcomes (Figure 1.3a). As described below, many tools and methods, mainly native-centric but some considering also the transition and denatured states, have been developed that may improve designs and increase the likelihood of obtaining a soluble and well-folded protein.

Using native-centric energy functions

Considerable progress both in whole sequence design and in subsequent optimization have been achieved using atomistic energy functions coupled with extensive sidechain/backbone configurational searching. In a relatively computationally tractable case, the backbone is kept fixed, limiting the search to only side-chain degrees of freedom. While successful designs have been realized using this approach [17, 44], the resulting sequences are frequently very similar to natural ones. Often, the ability of sequence design algorithms to recapitulate natural sequence statistics for a given topology is used as a scoring metric [88]. Such natural sequence statistics can be used directly as components of energy functions in design. For instance, Mitra *et al.* combined sequence information with the empirical FoldX energy function [89] to computationally redesign 243 proteins; 5 were tested experimentally of which ∼3 appeared by NMR to be quite well-folded [90]. Protein design has also been accomplished using only statistical energy functions as in the recent redesign of four natural targets [91]. Statistical energy functions, empirical or physical energy functions (like FoldX), and machine learning approaches have been applied as tools for stabilizing protein native structure, with mixed results [92, 93]; it is widely thought that more accurate force fields are needed to improve design outcomes [34, 94, 95]. Nevertheless, the abundance of computational tools for improving protein stability offers considerable scope for optimizing native structure (Figure 1.3b).

That design approaches employing a fixed backbone, as mentioned above, typically produce sequences with appreciable identity to natural proteins of the same topology, suggests the range of sequence innovation is limited. Varying the degrees of freedom of both the backbone and side chain, on the other hand, may greatly expand the range of designable targets, as even small (1-2 Å) perturbations of the backbone enabled the exploration of *de novo* sequences with low identity to natural counterparts or templates

[23, 38, 46]. Flexible backbone design can nevertheless recover natural sequences, and also recapitulate the covariation between amino acids at different positions [96]. The benefits of flexible backbone design, however, come at the cost of a more difficult optimization problem. This cost can be reduced by taking advantage of modularity and symmetry, as described for the largest completely *de novo* globular protein design to date (∼200 amino acid, see “Globular folds with internal symmetry”) [21] (Figure 1.2d). This design was also aided by previously developed rational rules for the design of backbone templates compatible with well-funnelled energy landscapes, which demonstrated impressive success for various αβ topologies [39, 21] (Figure 1.2h). Collectively, there has been much progress in applying native-centric approaches both for optimization of existing native structure and *de novo* design.

Beyond native-centric design, using coarse-grained simulations

Coarse-grained simulations are being used increasingly to move beyond the native state and analyze the energy landscape of designed proteins. These simulations have indicated less cooperative folding of the *de novo* designed Top7 topology compared to natural proteins, which may be a consequence of non-native interactions [97] and/or an imbalance of local and long-range interactions [98]. Similarly, other designed proteins may suffer from complex/non-cooperative folding kinetics [97, 99]. Coarse-grained simulations showed that designed surface electrostatic interactions in various proteins may reduce frustration and improve both equilibrium stability and folding kinetics [100] (Figure 1.3c). In contrast, non-native electrostatic interactions markedly slowed the folding of another designed protein in all atom molecular dynamics (MD) simulations [101]. In general, coarse-grained simulations have shown that the folding energy landscape is modulated in complex and quite often detrimental ways by functional features [102]; yet, foldability can be critical for achieving function [54]. Also, slower unfolding kinetics can protect proteins from degradation, modification, and aggregation, even if their thermodynamic stability is low; simulations illuminated how kinetic stability may be predicted and enhanced by increasing the proportion of contacts between residues distant in sequence [70] (Figure 1.3d). Thus, simulations are providing valuable insights into mechanistic details of folding, and so have potential to be a valuable tool to improve future designs and to assess the impact of designing functional features into idealized but function-less scaffolds [21, 39, 46].

Avoiding unintended oligomerization

A critical area for further development is controlling the population of the most stable folded protein structure relative to alternative conformations, e.g. oligomers or aggregates, or functional states. Many current designs fail to express a soluble protein or the intended monomeric form, often forming oligomeric species or specific domain-swapped dimers [22, 23, 39, 45, 47]. To address the common and specific problem of domain-swapped dimers, which may arise from highly native-centric design approaches, Mou *et al.* used atomistic MD to redesign a domain-swapped dimer into the intended monomer [103] (Figure 1.3e). To combat the tendency of forcefields used in design to generate hydrophobic patches on the protein surface — leading to unwanted oligomerization — forcefields have been re-parameterized to penalize such patches; this has improved the solubility of designs [104] (Figure 1.3f). Similarly, the design of high net-charge surfaces has resulted in increased protein solubility [105]. Aberrant oligomerization is a prevalent problem in design and these approaches may be widely applicable moving forward.

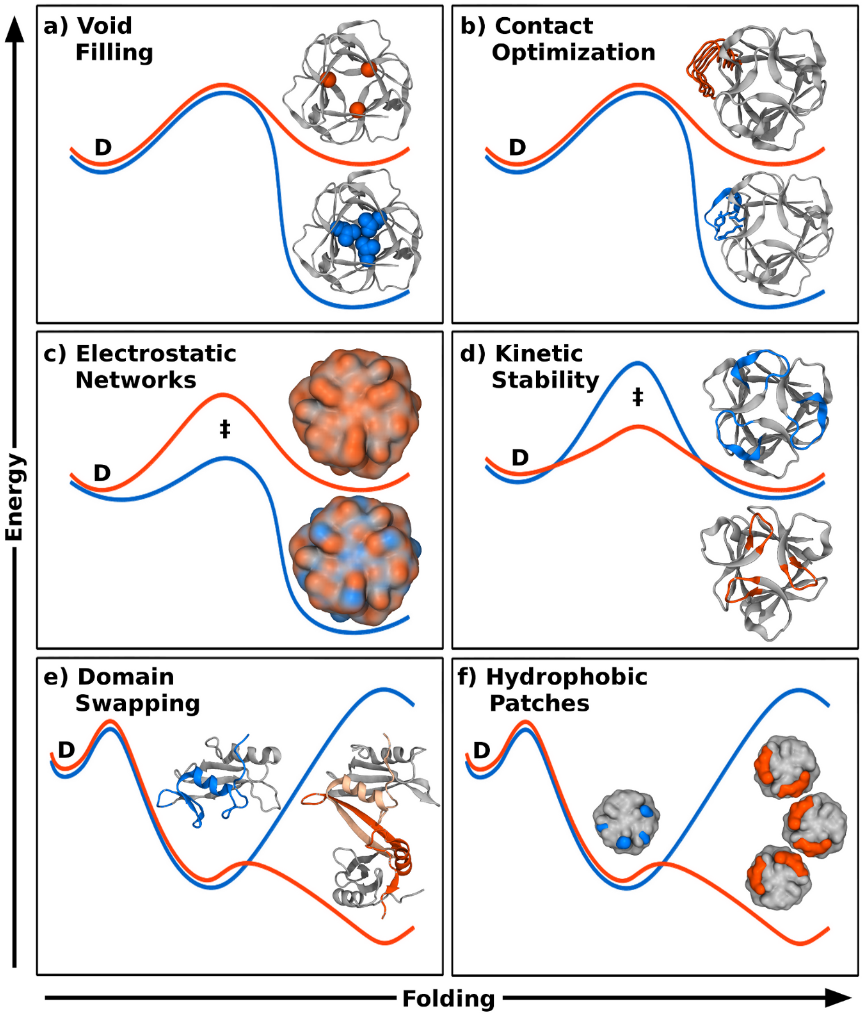


Figure 1.3: Optimizing designs. Energy profiles for initial (orange lines) and optimized (blue lines) protein designs; in each panel the folded state is at the right (and shown as structures) relative to the denatured state (D) at the left, separated by the transition state (‡). Optimization of features (blue parts of structures) of the initial designed or natural proteins (orange) can be designed by stabilizing the native state using CPD to (a) improve packing efficiency to eliminate voids (residues shown in space filling representation) [87], or (b) generally improve such properties as: polar contacts, sterics and backbone angles, among others, to improve energetics or reduce unwanted flexibility. Coarse-grained simulations of the entire energy landscape may also be used to (c) optimize electrostatic interactions [100] or (d) modulate topological complexity to control kinetic stability by tuning the energy barrier for unfolding [70]. Atomistic molecular dynamics can be employed to (e) eliminate unwanted oligomerization caused by local opening/domain swapping [103]. Aggregation of (f) exposed hydrophobic patches (orange) on designed surfaces may be eliminated (blue) by adding penalizing parameters to existing forcefields [104].

1.3.4 Conclusions

In recent years, designs of new sequences that stably adopt the intended natural, idealized, and even *de novo* topologies have been reported with increasing frequency. Ever-growing natural sequence and structure databases as well as increasing functional annotation [49] provide a valuable resource for further developing tools and methods to address outstanding challenges in achieving fully cooperative folding and incorporating function. Numerous examples have demonstrated that topological modularity and repetition/symmetry, a hallmark of many natural proteins, can simplify developing foldable and functional sequences by constraining the search space for computational design tools.

Despite the many successes, a substantial proportion of design attempts still fail [22, 23, 39, 47, 76, 77], and reliable discrimination of designs that will be successful or fail is not possible. While increasing the accuracy of forcefields may well improve results, typically native-centric approaches leave open the possibility that failed designs are unable to fold properly. However, increasing power in computation is opening pathways to explore the folding landscape during design, which may further improve outcomes [70, 100, 103].

Recent progress in developing increasingly accessible and sophisticated design approaches that make use of natural sequences, modularity, symmetry and coarse-grained as well as atomistic CPD have significantly advanced the outcomes and the scope protein topology design. Further work is needed to elucidate the detailed molecular basis for the specific characteristics of individual designed proteins. Such knowledge in combination with likely continuing methodological advances can be expected to drive the design of novel and useful proteins into a realm with a high confidence of success.

1.4 Molecular dynamics of ligand binding

While the design of protein structure as detailed in the previous section is a critical first step in realizing the full biomedical and industrial possibilities of proteins, a scaffold alone serves little practical value. Other than as a food source. In order to have function, a protein must first be capable of binding to something. Therefore, understanding how to rationally engineer binding specificity is paramount.

Correctly modelling protein-ligand binding is extremely difficult. This difficulty arises due to a combination of numerous degrees of freedom with a rugged energy landscape. Protein structures are highly dynamic, undergoing “breathing motions” on the same timescale that binding occurs [106], and even beyond the overall structure, sidechains themselves have considerable flexibility [107]. Thus, even a shallow binding groove on a protein surface can adopt a plethora of micro-configurations, without considering those of the ligand itself. Moreover, van der Waals forces, which account for a substantial fraction of the total energy of binding have a narrow energy minimum and any overlap between the van der Waals radii of contacting atoms, is highly unfavourable. These factors, and the vast ensemble of structures available to the binding partners, combine to generate an extremely large and rugged energy landscape when modelling or designing protein-ligand binding. If the approximate location of the binding site is not known and/or the ligand has appreciable flexibility or size (a flexible organic molecule like a carbohydrate, a small peptide, or even another protein) then the problem becomes yet more complex.

In chapter 7, the problem of protein-ligand binding is tackled using molecular dynamics (MD). MD has been used extensively over the last several decades, not only to tackle problems in protein-ligand binding [108], but also: ion transport through protein channels [109], protein folding [110, 111], protein aggregation [112, 113, 114], catalysis [115], and many others. Introducing the topic would be a volume in of itself, but a specific introduction of the methods used can be found in Chapter 7.

Chapter 2

Design of ThreeFoil *via* Repetition of a Subdomain Module

2.1 Context

This chapter is a pre-print version of “Modular Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein”, published in Structure in 2012 [44]. The work expands on an existing evolutionary hypothesis that proteins with internal structural symmetry emerged from repetition and fusion of smaller subdomain fragments or modules. The hypothesis was experimentally validated by reconstructing a protein, named ThreeFoil, which plausibly existed immediately after a series of repetition and fusion events that would have amplified the founder subdomain into a threefold symmetric globular protein. Additionally, an analysis of existing structures and sequences found that these repetition and fusions events are ongoing, at least in the case of the β-trefoil fold.

The original protein design work (Figure 2.4) and initial characterization (Figure 2.7, Figures 2.10 and 2.11) were done as part of my Masters degree. Andrew C. Doxey performed the bioinformatics analyses (Table 2.1, Figure 2.3, Figure 2.9, Table 2.3). Yuri D. Lobsanov collected X-ray diffraction data on crystals of ThreeFoil to solve the 3D structure (Table 2.2, PDB: 3PG0). During my PhD, I performed structural analyses to improve the sequence alignments and independently validated our hypothesis (Figures 2.1 and 2.2), grew the aforementioned crystals and analyzed the structure (Figure 2.5), arranged for and interpreted glycan micro-array data (Figure 2.6), and wrote the manuscript in collaboration with the other authors.

2.2 Summary

The high frequency of internal structural symmetry in common protein folds is presumed to reflect their evolutionary origins from the repetition and fusion of ancient peptide modules, but little is known about the primary sequence and physical determinants of this process. Unexpectedly, a sequence and structural analysis of symmetric subdomain modules within an abundant and ancient globular fold, the β-trefoil, reveals that modular evolution is not simply a relic of the ancient past, but is an ongoing and recurring mechanism for regenerating symmetry, having occurred independently in numerous existing β-trefoil proteins. We performed a computational reconstruction of a β-trefoil subdomain module and repeated it to form a newly three-fold symmetric globular protein, ThreeFoil. In addition to its near perfect structural identity between symmetric modules, ThreeFoil is highly soluble, performs multivalent carbohydrate binding, and has remarkably high thermal stability. These findings have far-reaching implications for understanding the evolution and design of proteins via subdomain modules.

2.3 Introduction

Internal structural symmetry is observed very frequently in common protein folds [48] and is thought to have arisen from the ancient evolution of these folds via the repetition and fusion of smaller peptide modules [40, 116]. The well-established occurrence of sequence duplication and fusion events in protein evolution [48, 117] supports the structural evidence for this evolutionary mechanism. However, in modern globular proteins, symmetry at the primary sequence level is typically relatively low to undetectable, owing to sequence divergence [40, 116]. This represents a challenge for understanding the origins and molecular determinants of symmetric protein evolution. Elucidating how symmetric protein structures can be constructed from a set of basic "building blocks" or subdomain modules has far-reaching implications not only for understanding evolution, but also for rational protein design.

Seminal studies on $(\beta\alpha)_8$-barrel proteins have provided experimental proof of principle for the evolution of symmetric globular folds via the repetition of subdomain modules [118, 53, 119]. Sterner, Hocker, and colleagues identified sequence and structural evidence for the evolution of this fold from a $(\beta\alpha)_4$-half-barrel ancestor [120]. By fusing two identical copies of a half-barrel and stabilizing the resulting protein using a combination of rationally designed mutations and mutations selected from a library of variants, they obtained a stable and symmetric structure, although it was lacking in function [119, 52]. As protein design experiments often fail to produce the intended structure or properties [118, 17, 121], and data for other globular symmetric folds is limited, additional investigations are needed. The recent explosive growth in the availability of protein sequences and structures from genomics initiatives combined with new tools for reconstructing and designing proteins have set the stage for such investigations.

The focus of this study is the internally symmetric β-trefoil structure, an ancient fold adopted by many proteins with a great diversity of sequences and ligand-binding functions [122, 123]. β-trefoils currently include at least 14 families according to Pfam [124], such as the carbohydrate-binding ricin and agglutinin toxins, actin-bundling proteins, the fibroblast growth factor (FGF) and interleukin-1 cytokines, STI-like protease inhibitors, and LAG-1 DNA-binding proteins. The β-trefoil fold displays three-fold internal structural symmetry (Figure 2.1A), and internal sequence similarities have been noted in some families, such as the multivalent sugar-binding ricins and actin-bundling proteins [122]. Each of the three subdomain modules is composed of four β-strands, with two strands from each module collectively forming a six-stranded β-barrel and the remaining two from each module together forming a β-hairpin triplet that caps one end of the barrel. Previous sequence analyses have suggested that modern β-trefoil proteins share a common homotrimer ancestor of identical subdomain modules [122, 123, 125, 126]. A recent parallel study to the work herein demonstrated the feasibility of this evolutionary model by constructing both homotrimer and fused three-fold symmetric β-trefoil structures, developed from a cytokine FGF template using rational design and library screening, in which protein function was lost [127, 79]. This, together with a wealth of previous structural and folding studies on β-trefoil proteins [123, 128, 129, 130], makes them an attractive candidate for examining modular evolution.

We report herein a large-scale sequence clustering analysis of β-trefoils. Based on previous analyses, we expected to find evidence for a single ancient homotrimer and triplication event [122, 123, 125, 126]. To our surprise, our analysis revealed that the repetition and fusion of subdomain modules to form new symmetric β-trefoils is an ongoing and recurring process that has occurred numerous times. We then reconstructed a completely three-fold symmetric β-trefoil sequence by using consensus sequence and protein design based on a carbohydrate-binding ricin sequence identified from the clustering results. The designed protein, ThreeFoil, forms a structure whose subdomain modules are essentially identical, and further more, it exhibits extremely high thermal stability, as well as functional multivalent carbohydrate-binding properties. The single subdomain module, OneFoil, is poorly structured, however. Consequently, incorporating symmetry may be attractive both in evolution and in the design of multivalent binding proteins.

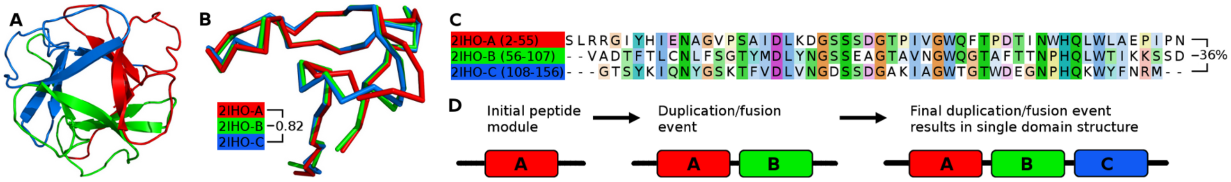

A
B
2IHO-A
2IHO-B
2IHO-C
0.82
C
2IHO-A (2-55) SLRRGIYHIENAGVPSAIDLKDGSSSDGTPIVGWQFTPDTINWHQLWLAEPIPN
2IHO-B (56-107) --VADTFTLCNLFSGTYMDLYNGSSEAGTAVNGWQGTAFTTNPHQLWTIKKSSD
2IHO-C (108-156) ---GTSYKIQNYGSKTFVDLVNGDSSDGAKIAGWTGTWDEGNPHQKWYFNRM--
36%
D
Initial peptide module
Duplication/fusion event
Final duplication/fusion event results in single domain structure
A
A
B
A
B
C

Figure 2.1: Internal symmetry in the β-trefoil fold. (A) Structure of a typical β-trefoil domain (the *Marasmius oreades* mushroom lectin, PDB ID 2IHO [131]), looking down the threefold symmetry axis. Subdomains are colored red, green, and blue from N- to C-terminus. (B) Structural alignment of the Cα trace of subdomains, aligned using SSM [132] and showing the average structural identity defined by SSM's Q-score parameter. (C) The sequence alignment of 2IHO subdomains. The average sequence identity between pairs of aligned subdomains as determined using MUSCLE [133] is 36%. (D) One possible model of β-trefoil evolution from a single subdomain. All structure images were rendered using PyMol (http://www.pymol.org/).

2.4 Results

2.4.1 Sequence analysis of β-trefoil subdomains reveals recurring modular evolution

We analyzed the evolutionary relationships among β-trefoil subdomain modules by constructing a dataset of subdomain modules and clustering these according to sequence similarity. First, a dataset of 1167 nonredundant sequences annotated as β-trefoils was obtained using the Conserved Domain Database from the National Center for Biotechnology Information [134]. This set included members of 11 β-trefoil families, each with a representative of known structure (Table 2.1). Through alignment to representative structures, each β-trefoil sequence was subdivided into three β-β-β-loop-β subdomains, the putative building block of the β-trefoil fold [123] (Figure 2.1B,C). In order to assess the evolutionary relationships between the subdomains, they were clustered by sequence similarity, where each subdomain pair with E $<$ 1e-04 was connected.

Remarkably, we found a pattern of greater similarity between subdomains within a given β-trefoil sequence than between subdomains from different β-trefoils. Together these findings reveal ongoing evolution in which a distinct single-subdomain module was repeated to form a new symmetric protein. The predominant accepted model of protein domain evolution is the duplication and divergence of whole domains [135]. According to this model, a given β-trefoil module should be most similar to the same module in a closely related sequence. Indeed, this mode of evolution is observed for the majority of subdomains, as is illustrated for a pair of proteins (Figure 2.2A), and for the entire dataset of sequences (representative clusters in Figure 2.3B; all clusters in Figure 2.9C). Strikingly, however, there are also multiple examples where each β-trefoil subdomain module is most similar to the other two modules within the same protein sequence, and less similar to the modules of other closely related β-trefoils. We illustrate this pattern for a pair of proteins (Figure 2.2B) and for the entire dataset of sequences, where we identified nine cases of subdomain module repetition through our clustering analysis, in the ricin, AbfB, and fascin families (representative clusters in Figure 2.3C; all clusters in Figure 2.9B; sequence alignments for representatives of each cluster in Table 2.3). To more sensitively detect subdomain-repetition events, including those occurring within a cluster, we performed a phylogenetic analysis of the subdomains showing the highest internal symmetry and identified nine additional (i.e., 18 total) distinct subdomain-repetition events (Figure 2.9D). These repetition events occurred most prominently within the ricin family, which included the eight sequences with greater similarity between subdomains than with any other subdomains (Figure 2.9D,E). This pattern of greater internal than external similarity demonstrates ongoing evolution of new β-trefoil folds by repetition of subdomain modules. The size of the clusters (Figure 2.3C) and the interrelationship of subdomains (Figure 2.9D) shows that such subdomain-repetition events may be preceded or followed by whole-domain duplication. A similar process of subdomain repetition has been postulated for the nonglobular β-propeller fold [136], and may apply for $(\beta\alpha)_8$-barrels [53] and many other internally symmetric protein folds (see Discussion).

Table 2.1: Dataset construction and calculated sequence symmetries

Family	CDD IDs[a]	Extracted Domain Sequences[b]	Domains After Filtering[c]	Average Sequence Symmetry[d]	Representative Structure Used for Alignment
AbfB	68828, pfam05270	24	15	17.7	1W3D
Agglutinin	70918, pfam07468	14	5	9.0	1JLX
CD Toxin	80015, pfam03498	69	28	9.2	1SR4
Fascin	29332, pfam06268, cd00257 87053	413	129	13.0	1DFC
FGF	28940, pfam00167, cd00058 47749, smart00442 84576	775	140	10.3	1NUN
IL1	28984, pfam00340, cd00100 64217	362	86	8.2	1MD6
STI/Kunitz	29140, pfam00197, cd00178 84601	452	89	7.7	1WBA
LAG1	72686, pfam09270	31	18	7.0	1TTU
MIR	86128, pfam02815	1267	65	10.3	1T9F
Ricin	29101, pfam00652, cd00161 47764, smart00458 84930	1604	518	14.3	1QXM
Toxin R Bind C	87408, pfam07951	89	15	7.3	3BTA

[a] See ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/cdd versions for conserved domain model accessions and version information

[b] Sequences of "all related families" for each CDD ID were retrieved from the NCBI

[c] Filtering involved removal of redundancy and partial sequences (see Methods)

[d] The mean pairwise percent identity between the three repeats in each domain was calculated. The average sequence symmetry is the mean of this value for all domains in the family

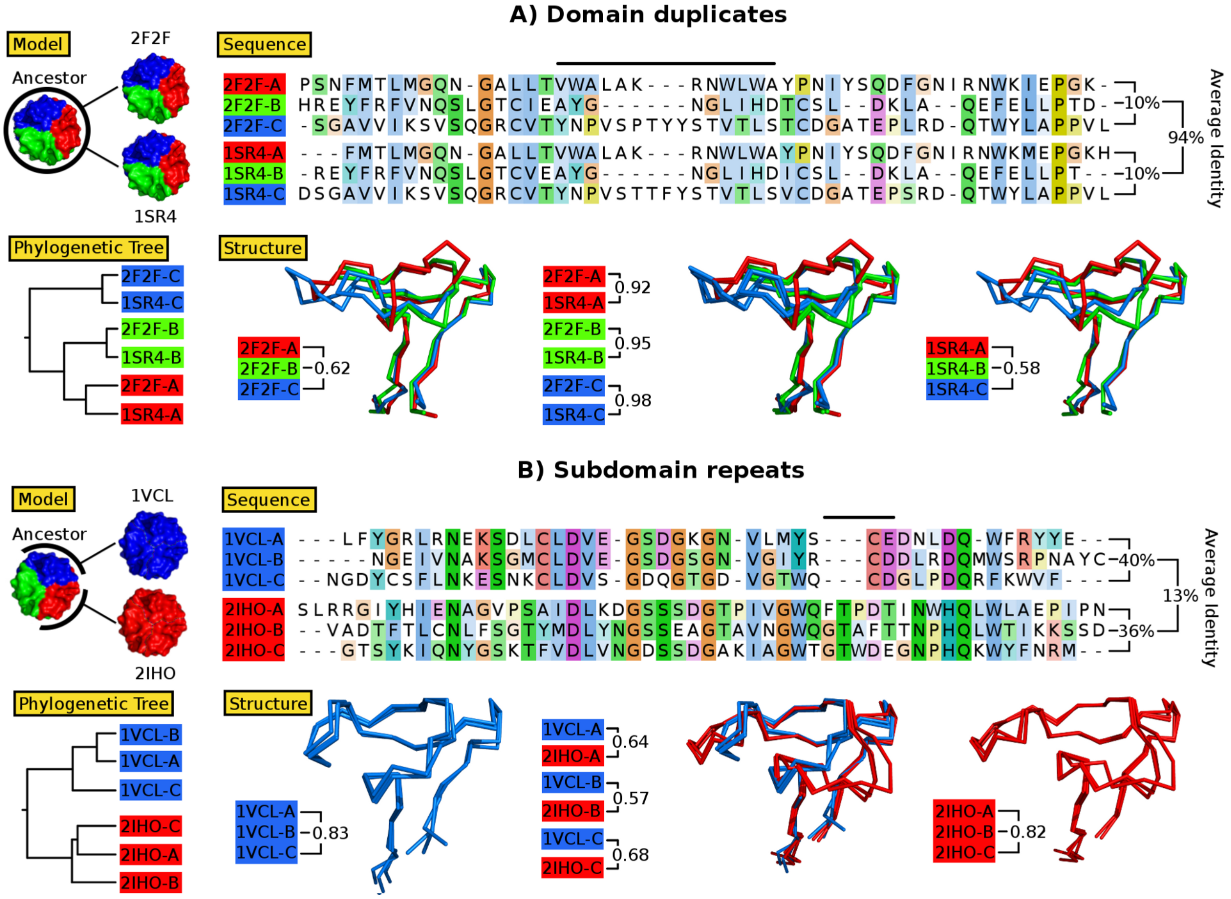

A) Domain duplicates
Model
2F2F
Ancestor
1SR4
Sequence
2F2F-A PSNFMTLMGQN-GALLTVWALAK---RNWLWAYPNIYSQDFGNIRNWKIEPGK-
2F2F-B HREYFRFVNQSLGTCIEAYG------NGLIHDTCSL--DKLA--QEFELLPTD-
2F2F-C -SGAVVIKSVSQGRCVTYNPVSPTYYSTVTLSTCDGATEPLRD-QTWYLAPPVL
1SR4-A ---FMTLMGQN-GALLTVWALAK---RNWLWAYPNIYSQDFGNIRNWKMEPGKH
1SR4-B -REYFRFVNQSLGTCVEAYG------NGLIHDICSL--DKLA--QEFELLPT--
1SR4-C DSGAVVIKSVSQGRCVTYNPVSTTFYSTVTLSVCDGATEPSRD-QTWYLAPPVL
10%
10%
94%
Average Identity
Phylogenetic Tree
2F2F-C
1SR4-C
2F2F-B
1SR4-B
2F2F-A
1SR4-A
Structure
2F2F-A
2F2F-B
2F2F-C
0.62
2F2F-A
1SR4-A
0.92
2F2F-B
1SR4-B
0.95
2F2F-C
1SR4-C
0.98
1SR4-A
1SR4-B
1SR4-C
0.58
B) Subdomain repeats
Model
1VCL
Ancestor
2IHO
Sequence
1VCL-A ---LFYGRLRNEKSDLCLDVE-GSDGKGN-VLMYS---CEDNLDQ-WFRYYE--
1VCL-B -----NGEIVNAKSGMCLDVE-GSDGSGN-VGIYR---CDDLRDQMWSRPNAYC
1VCL-C --NGDYCSFLNKESNKCLDVS-GDQGTGD-VGTWQ---CDGLPDQRFKWVF---
2IHO-A SLRRGIYHIENAGVPSAIDLKDGSSSDGTPIVGWQFTPDTINWHQLWLAEPIPN
2IHO-B --VADTFTLCNLFSGTYMDLYNGSSEAGTAVNGWQGTAFTTNPHQLWTIKKSSD
2IHO-C ---GTSYKIQNYGSKTFVDLVNGDSSDGAKIAGWTGTWDEGNPHQKWYFNRM--
40%
36%
13%
Average Identity
Phylogenetic Tree
1VCL-B
1VCL-A
1VCL-C
2IHO-C
2IHO-A
2IHO-B
Structure
1VCL-A
1VCL-B
1VCL-C
0.83
1VCL-A
2IHO-A
0.64
1VCL-B
2IHO-B
0.57
1VCL-C
2IHO-C
0.68
2IHO-A
2IHO-B
2IHO-C
0.82

Figure 2.2: evolution by whole-domain duplication *versus* subdomain repetition. β-trefoils are labelled with their PDB IDs, and different subdomains are colored as in Figure 2.1. Evolution by whole-domain duplication (A) and subdomain module repetition mechanisms (B). For each mechanism, a set of representative amino acid sequences given in single letter code illustrates the respective mode of evolution. In (A) and (B), a structural representation of each evolutionary mode is shown with space-filled β-trefoil structures (Model), illustrating the putative evolutionary path, which is supported by the phylogenetic tree inferred from the sequence alignment (Sequence; black bars highlight key regions) using Phylip (http://evolution.genetics.washington.edu/phylip.html) and MUSCLE [133], respectively. In addition, the same pattern is seen in structure alignments (Structure) made using SSM [132]. Sequence identities were given by MUSCLE and structural similarities were defined by the Q-Score parameter in SSM (with a score of 1.0 representing identical C_α traces).

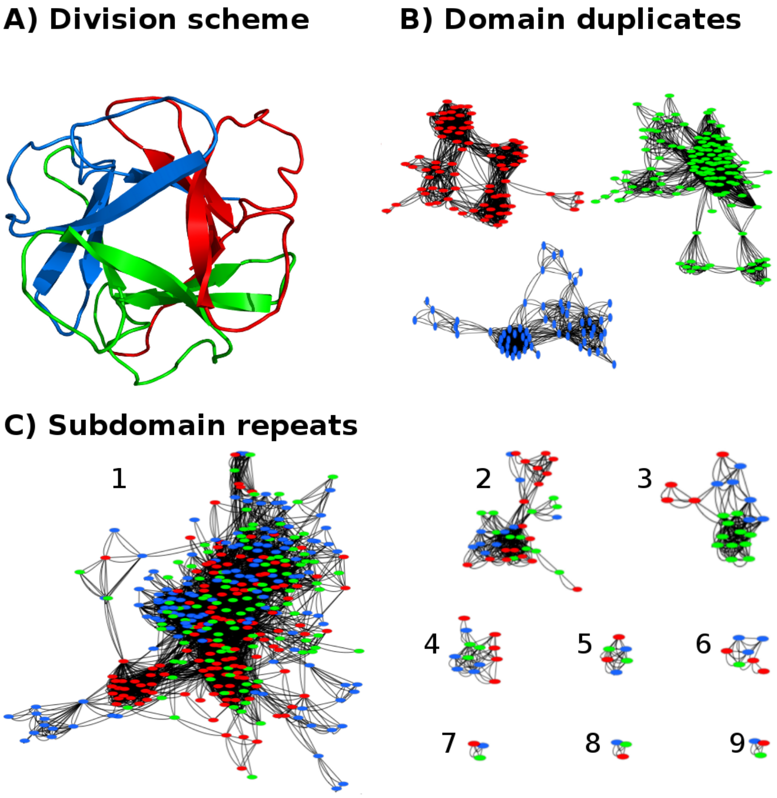


Figure 2.3: representative sequence clusters of β-trefoil subdomain modules show occurrences of both whole-domain duplication and subdomain repetition. Individual subdomain modules represented by an oval are colored as in Figure 2.1 and clustered according to sequence similarity, as described in the Methods and Results sections. (A) Division scheme for splitting of β-trefoil domains into three constituent symmetrical subdomain modules. (B) Representative clusters demonstrating whole-domain duplication and divergence resulting in subdomains that are most closely related to the corresponding subdomain in homologous sequences. (C) Clusters demonstrating evolution *via* subdomain repetition. Internal subdomains are more closely related to one another than to extant subdomains. Clusters are numbered as in Table 2.3, and all clusters can be seen in Figure 2.9C. In addition, a phylogenetic tree and heat map of the most internally symmetric subdomains are shown in Figure 2.9D, with a boxplot of internal symmetry by sequence family in Figure 2.9E.

2.4.2 Reconstructing a progenitor subdomain module through sequence analyses and computational design

We tested the physical feasibility of evolution *via* subdomain repetition and fusion by using a combination of sequence analysis and computational design to reconstruct a β-trefoil consisting of three identical subdomain modules. We reasoned that the natural β-trefoil sequence with the highest internal sequence symmetry identified in the clustering analysis would be a good starting point. This was a member of the ricin family annotated as the carbohydrate binding module of a glycosidase from the halophilic red archaeon *Haloarcula marismortui* (NCBI accession no. AAV45265), which has 55% amino acid identity between the three modules.

In order to reconstruct a completely symmetrical β-trefoil, ThreeFoil, three steps were used to incorporate information from the template sequence, homologous sequences, and rational protein design. In the first step, the template sequence was split into its three constituent subdomain modules (Figure 2.4, Template), and those residues conserved in all three modules were fixed (Figure 2.4, Step 1); this left 21 of 47 positions undefined. In the second step, a small set of 13 highly homologous sequences were identified and aligned with the template sequence and split into their corresponding subdomain modules, and the residue frequency was calculated at each position (Figure 2.4, Homology; see Figure 2.10 for homologous subdomain module alignments). The residue frequency at each position was averaged between the homologous sequences and the template, and any residue with an average frequency >0.5 (50%) was incorporated into the reconstructed sequence (Figure 2.4, Step 2). This left 16 positions undefined.

The third step of reconstruction made use of computational design in the form of Rosetta Design [17]. Allowing only the 16 undefined positions to vary, Rosetta Design generated a set of 10,000 energetically favorable sequences, and the residue frequency at each position was calculated (Figure 2.4, Rosetta). Two points of concern were limitations in the successful design of all-β proteins using Rosetta Design [17, 137] and the low sequence conservation at some positions. To address this, the overall residue frequency at each position was calculated by equally weighting the frequency of residues in the template, in the homologous sequences, and from Rosetta Design, with the most frequent residue at a given position being incorporated into the final reconstructed sequence (Figure 2.4, Step 3). This approach allowed for inclusion of residues important for function and stability based on consensus information from the template and homologous sequences [67]; at the same time, it allowed energetically favorable residues identified by Rosetta Design to be incorporated. In this respect, it is reassuring to note that the residues most frequently identified by Rosetta Design were well represented in the homologous sequences.

The final reconstructed subdomain module sequence was expressed as a single module, OneFoil, and as a fused three-fold repeat, ThreeFoil. The proteins were characterized as described in the following section.

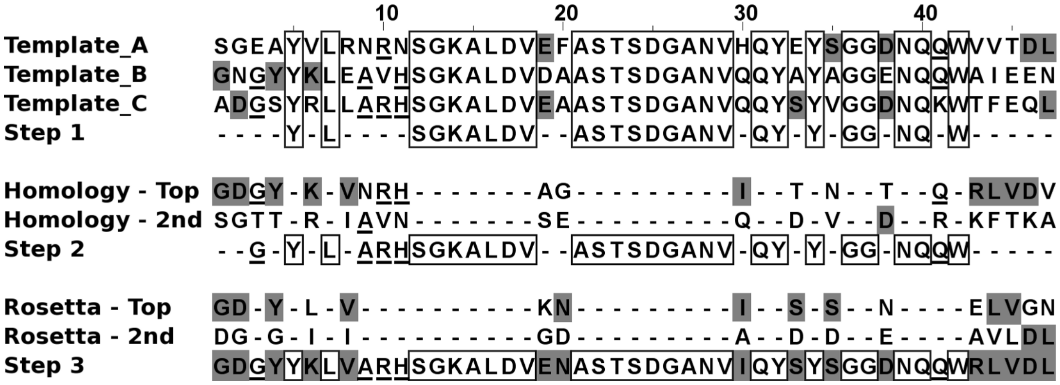

10
20
30
40
Template_A SGEAYVLRNRNSGKALDVEFASTSDGANVHQYEYSGGDNQQWVVTDL
Template_B GNGYYKLEAVHSGKALDVDAASTSDGANVQQYAYAGGENQQWAIEEN
Template_C ADGSYRLLARHSGKALDVEAASTSDGANVQQYSYVGGDNQKWTFEQL
Step 1 ----Y-L----SGKALDV--ASTSDGANV-QY-Y-GG-NQ-W-----
Homology - Top GDGY-K-VNRH-------AG---------I--T-N--T--Q-RLVDV
Homology - 2nd SGTT-R-IAVN-------SE---------Q--D-V--D--R-KFTKA
Step 2 --G-Y-L-ARHSGKALDV--ASTSDGANV-QY-Y-GG-NQQW-----
Rosetta - Top GD-Y-L-V----------KN---------I--S-S--N----ELVGN
Rosetta - 2nd DG-G-I-I----------GD---------A--D-D--E----AVLDL
Step 3 GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL

Figure 2.4: Reconstruction of a three-fold symmetric sequence, ThreeFoil. The template sequence is shown split into its three subdomain modules (Template A to Template C), with conserved residues used to reconstruct the partial identity of the putative progenitor subdomain module (Step 1, boxed residues). The frequency of amino acids at each remaining position in a set of homologous sequences (see Figure 2.9) (most frequently seen amino acids in Homology [top], second most frequent in Homology [2nd]) were used together with the template sequence for further reconstruction (Step 2, underlined residues). The frequency of amino acids at the remaining positions, as predicted in a set of low-energy Rosetta Design models (most frequently predicted amino acids in Rosetta [Top], second most frequent in Rosetta [2nd]), was used with the template and homologous sequence data to complete reconstruction (Step 3, gray highlighting). For more details, see Results and Methods.

2.4.3 ThreeFoil has near-perfect three-fold structural and ligand-binding symmetry

The structure of ThreeFoil was determined by X-ray crystallography to a resolution of 1.62 Å , and refined to high quality (see Table 2.2 for refinement statistics; structure deposited as PDB code 3PG0). The structure exhibits exceptionally high symmetry, as evidenced by a very low backbone RMSD of only 0.2 Å between subdomain modules (Figures 2.5A,B,C). The symmetry of ThreeFoil is also apparent in its binding of ligands, including galactose, a metal ion, and ordered water molecules. The template sequence from *Haloarcula marismortui* is annotated by BLAST [138] as a member of the ricin family of β-trefoils, which bind carbohydrates (often terminating in galactose) in a shallow pocket formed by the second and third β-strands and the long loop between strands 3 and 4 [139]. This pocket in ThreeFoil contains bound bis-tris from the crystallization buffer in all three symmetric units (Figure 2.5D). Bis-tris has been shown previously to occupy expected active or binding sites within a protein [140], and given its many hydroxyl groups, it likely mimics the natural carbohydrate ligand. The binding of D-galactose to ThreeFoil was measured *via* changes in the intrinsic protein fluorescence upon sugar binding (Figure 2.6A), giving a dissociation constant (K_d) of ∼1 mM, which is very similar to the measured K_d for D-galactose binding to one of the proteins used in the homology modelling of ThreeFoil [141]. In addition, the binding of ThreeFoil to a series of glycans was measured using a glycan array. The results clearly show that ThreeFoil's symmetry allows for multivalent binding, as seen in the pronounced improvement in binding from a single glycan chain to a multi-antennary one (Figure 2.6B).

In addition to binding carbohydrates, many β-trefoil structures have structurally conserved buried water molecules in each symmetrical unit [123]. ThreeFoil also binds three symmetrical buried water molecules, which make important bridging hydrogen bonds between strands 1, 2, and 4 of each symmetrical unit (Figure 2.5E). Finally, ThreeFoil binds a single metal ion along the three-fold axis of symmetry, which coordinates one backbone oxygen atom and one side-chain asparagine oxygen atom from each symmetric unit (Figure 2.5F) in an octahedral manner. The presence of a metal ion located on the axis of symmetry is very common for cyclically symmetric protein structures [142], and may point to a primordial role for metal ions in stabilizing symmetric structures.

Table 2.2: Data collection and refinement statistics

Data Collection	
Space group	$P4_32_12$
Cell dimensions	
$a,\ b,\ c$ (Å)	45.0, 45.0, 113.4
$\alpha\ \beta\ \gamma$ (°)	90.0, 90.0, 90.0
Resolution (Å)	1.62 (1.68-1.62)[a]
R_{merge} [b] (%)	0.068 (0.318)
Average I/σI	13.7 (3.5)
Completeness (%)	99.5 (96.0)
Redundancy	6.25 (4.29)
Refinement	
Resolution (Å)	1.62
No. measured reflections	97142
No. unique reflections	15533
R_{cryst}/R_{free} [c]	16.7/18.5
No. atoms	
Protein	1151
Ligand/ion	61
Water	115
Average *B*-factors (Å^2)	
Protein	14.7
Ligands[d]	
BTB 1,2,3;	10.8, 22.6, 34.4;
Glycerol 1,2;	38.5, 38.5;
Na^+ ion	10.6
Water	28.5
RMSDs	
Bond lengths (Å)	0.006

Bond angles (°)	1.07
Coordinate error (ML-based, Å)(8)	0.22
Ramachadran plot (%)	
Most favored	86
Allowed	14

[a] Values in parentheses are for last resolution shell

[b] $R_{merge} = \sum\sum|I(k)-<I>|/\sum I(k)$, where I(k) is the measured intensity for each symmetry related reflection and <I> is the mean intensity for the unique reflection. The summation is over all unique reflections

[c] $R_{cryst} = \sum|F_o| - |F_c|/\sum|F_o|$ and $R_{free} = \sum(|F_{os}| - |F_{cs}|)/\sum|F_{os}|$, where "s" refers to a subset of the data not used in the refinement, representing 7% of the total number of observations

[d] Ligand atoms BTB 1,2,3 and glycerol 1,2 refer to three Bis-Tris methane molecules of Bis-Tris Buffer (BTB) and two glycerol molecules of the cryoprotectant that were identified in the electron density and built into the structure

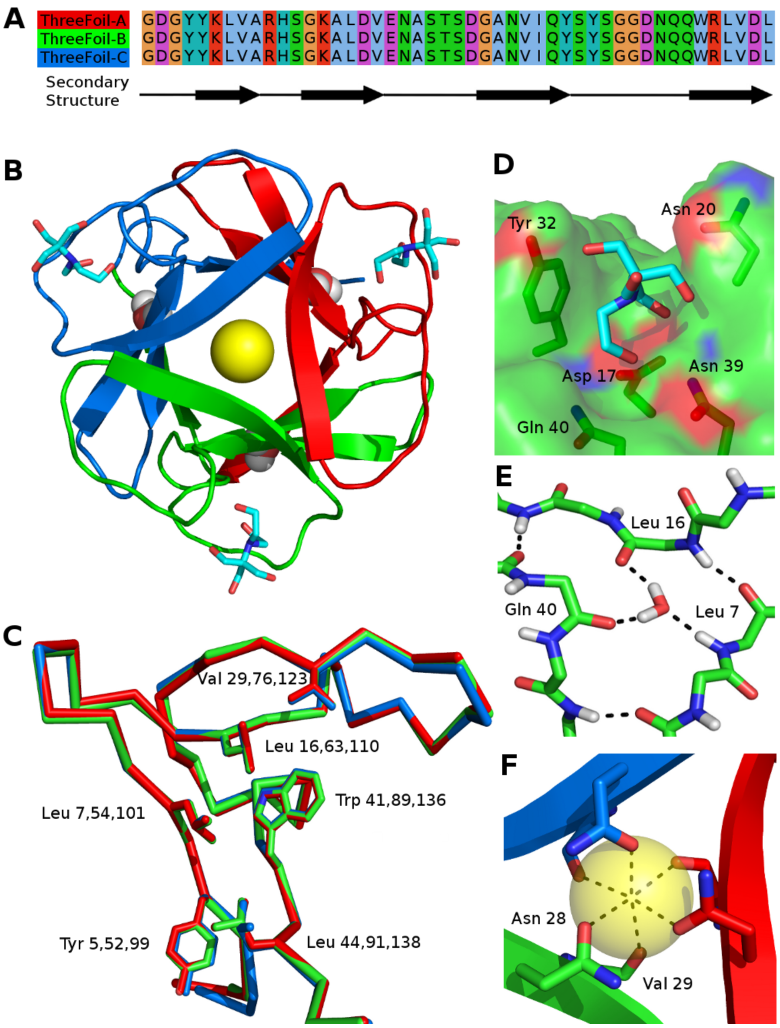

A
ThreeFoil-A GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
ThreeFoil-B GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
ThreeFoil-C GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL
Secondary Structure
B
C
Val 29,76,123
Leu 16,63,110
Trp 41,89,136
Leu 7,54,101
Tyr 5,52,99
Leu 44,91,138
D
Tyr 32
Asn 20
Asp 17
Asn 39
Gln 40
E
Leu 16
Gln 40
Leu 7
F
Asn 28
Val 29

Figure 2.5: Symmetric structure in ThreeFoil. (A) The three subdomain modules of ThreeFoil aligned with the secondary structure shown below. (B) A view of ThreeFoil along its three-fold symmetry axis, with subdomains indicated using the same colors as in Figure 2.1, bound bis-tris carbon atoms in cyan, bound waters as red (oxygen) and white (hydrogen) spheres, and the bound sodium as a yellow sphere. (C) Each subdomain module of ThreeFoil structurally aligned by C_α using SSM [132] shown as a C_α trace, with core hydrophobic sidechains shown as sticks. (D) Bis-tris bound to ThreeFoil in the shallow pocket that forms the carbohydrate binding site in related ricins. (E) The buried water molecule in each subdomain forms hydrogen bonds with three different β-strands. (F) Sodium binding site, showing the symmetric backbone and side-chain oxygen atoms involved in the octahedral coordination.

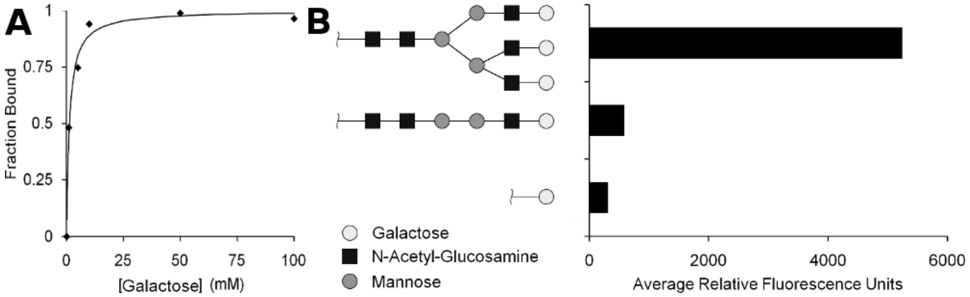


Figure 2.6: Multivalent glycan binding by ThreeFoil. (A) Galactose binding curve for ThreeFoil measured by fluorescence showing a dissociation constant (K_d) of 1 mM. (B) Binding results from a glycan array demonstrating that ThreeFoil has considerably improved binding to multi-antennary glycan structures (top structure as compared with bottom two).

2.4.4 ThreeFoil is well behaved in solution: monomeric, highly soluble, and extremely stable

Since many designed proteins have a tendency to misfold or aggregate [118, 17, 121], we further tested the success of the ThreeFoil design using a battery of biophysical measurements. These showed that ThreeFoil is a highly soluble monomer with high thermal stability. Static light scattering (SLS) (Figure 2.7A), dynamic light scattering (DLS), and size exclusion chromatography (SEC) (Figures S3A and S3B) showed that ThreeFoil is a highly soluble monomer in solution. ^{1}H-NMR spectroscopy showed that ThreeFoil is well folded and has a well defined structure (Figure 2.7B), with numerous downfield amide resonances, as expected for β-sheet structure, and upfield methyl resonances indicating a well packed hydrophobic core. In addition, ThreeFoil unfolds at a high temperature of ~94 °C by differential scanning calorimetry (DSC) (Figure 2.7C), further demonstrating its stability.

It is interesting that the single-peptide module used to generate ThreeFoil, termed OneFoil, is sufficiently stable for expression; however, it appears to be unfolded, based on NMR (Figure 2.7D) and fluorescence (Figure 2.9C). In contrast, fluorescence spectroscopy of ThreeFoil showed that aromatic residues undergo a very pronounced blue shift upon folding (Figure 2.9D), characteristic of burial in a solvent-inaccessible hydrophobic core [143]. This indicates a significant energetic penalty for forming the β-trefoil fold from multiple smaller chains, as has also been reported for other proteins [127, 79, 144]. This suggests the possibility that while the first symmetry-forming event for β-trefoils may have proceeded from a homotrimer [122, 123, 125, 126], the recurring symmetry-forming events highlighted by our analysis may proceed from a subdomain module within an existing whole domain, thereby avoiding the energetically penalized homotrimeric form and also suggesting an explanation for why no isolated subdomain sequences have been reported.

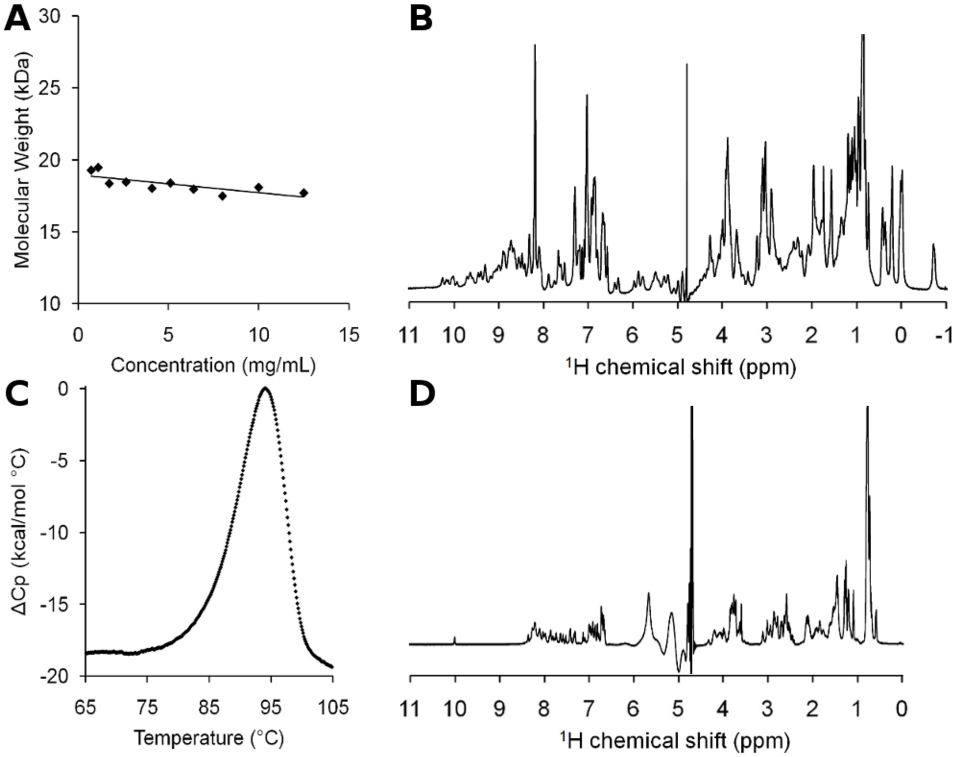


Figure 2.7: Biophysical characterization of ThreeFoil. (A) Molecular weight Debye plot [145] of SLS measurements, consistent with expected size of a ThreeFoil monomer (see also Figure 2.9). (B) ^{1}H-NMR spectrum of ThreeFoil in H_2O (containing 7% D_2O). The relatively sharp and well dispersed lines are indicative of a well folded monomeric structure. (C) DSC of ThreeFoil showing a large endothermic peak is typical of a cooperative thermal unfolding transition with a midpoint of ~94 °C. (D) ^{1}H-NMR spectrum of OneFoil in H_2O (containing 7% D_2O) with features typical of an unfolded protein (lack of amide resonances >8.5 ppm, and methyl resonances <1 ppm).

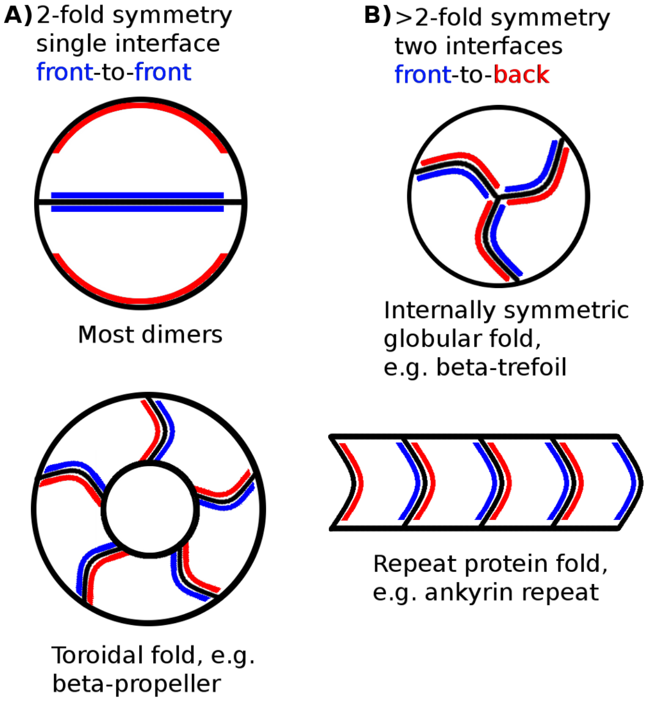


Figure 2.8: Packing of symmetric, repeat, and oligomeric proteins. Proteins with two-fold internal symmetry and dimers both tend to interact through a single interface, packing front to front (A). Proteins with more than two-fold symmetry or oligomers require two different interfaces and pack front to back (B).

2.5 Discussion

Our protein sequence analyses and design results provide exciting experimental support for ancient as well as ongoing evolution of globular proteins *via* the repetition of subdomain modules. The recurring symmetry-forming events in globular folds are a surprising discovery, challenging the common view that symmetric globular folds were generated only in the ancient past and subsequently diverged, losing internal sequence symmetry [40, 116, 122, 127]. Several groups have undertaken to make fully symmetric versions of common globular folds from different structural classes: two-fold symmetric four-helix bundles, two- and four-fold symmetric $(\beta\alpha)_8$-barrels [118, 53, 119, 144, 146], and now, in this study and a parallel study on FGF by the Blaber group [127, 79], two very different three-fold symmetric β-trefoils. Among these designed proteins, demonstration of successful design by biophysical and structural analyses has been reported only for a $(\beta\alpha)_8$-barrel [119] and for β-trefoils [127, 79].

The rational design and selection approach used by the Blaber group differs from the bioinformatics and rational design approach herein in that it uses multiple rounds of incorporating a few selected mutations to gradually increase symmetry, followed by screening for stability, ultimately resulting in a highly stable but nonfunctional protein. The primary sequences of the FGF design and ThreeFoil, which are based on proteins from different β-trefoil superfamilies (cytokines and ricin toxins, respectively), are well below the twilight zone and into the midnight zone of similarity (only ∼15% identity) [147]. Thus, the results reported here may illustrate how common folds can persist in evolution due to their compatibility with highly diverse sequences [135]. Furthermore, common symmetric folds may be highly populated because they are generated repeatedly.

Our results for the β-trefoil fold have broad implications for understanding evolution not only of symmetric globular proteins but of other types of protein structures containing repeated subdomain modules, in particular, elongated repeat proteins and toroidal proteins (such as β-propellers), as well as oligomeric proteins. Consideration of the types of interfaces between modules in these proteins reveals key structural relationships (Figure 2.8). In proteins containing two repeated modules, or homodimers, a single, typically symmetric, interface is formed between the repeated structural elements. In all proteins with more than two repeats, at least two distinct interface surfaces are formed [142]. For example, in β-trefoils each subdomain module packs front to back against the other two modules. Similarly, front-to-back packing of multiple ∼20- to 40-amino-acid modules is the basis for the structure of elongated repeat proteins and toroidal proteins. It is important to note, though, that repeat and toroidal proteins are fundamentally different from globular proteins, because their hydrophobic core lacks interactions between residues that are distant in the primary sequence [148, 149]. The front-to-back packing and globular arrangement of modules in β-trefoils is also similar to that frequently observed in compact homooligomers with more than two subunits and cyclic symmetry [142]. Thus, ThreeFoil shares structural characteristics with many other protein folds.

Common structural characteristics may underlie intriguing similarities in the stability and folding of β-trefoils, toroidal proteins, repeat proteins, and homooligomers. Consensus-sequence designed repeat proteins containing identical repeats have been found to have very high stability [148, 67]. For repeat proteins, stability increases with increasing number of identical repeats [148, 150]. Similarly, additional interfaces may also result in oligomers being generally more stable than monomers [151]. Additional entropic stabilization of symmetric globular proteins may be obtained by combining subdomain modules into a single chain. Such stabilization is suggested by the well structured ThreeFoil compared with the unstructured OneFoil, and by similar results for a symmetric FGF β-trefoil [127, 79] and a four-helix bundle protein [144]. Thus, combining identical modules into a single chain could favor folding in multiple ways. In general, the roles of structural symmetry in protein folding are not yet well understood, and proteins with completely symmetric tertiary structure and sequence, like ThreeFoil, are intriguing models for examining these roles further.

Internally symmetric structures may provide significant benefits for protein function. For instance, β-trefoils and toroidal, repeat, and oligomeric proteins appear to be particularly well suited for a wide range of binding functions, and they often use multiple repeats for multivalent binding of ligands, as seen with ThreeFoil and reported in many other cases [123, 139, 148, 152]. Thus, these protein structures may provide stable scaffolds for displaying a wide variety of loop structures for binding diverse ligands. It is now widely accepted that functional features such as binding sites can be a significant source of instability in proteins [153, 154], and symmetrical structures may be more stable [151], as well as more robust to mutations and therefore more designable [151, 155]. Together, this suggests that the repetition of structural modules in proteins may confer sufficient stability to accommodate destabilizing functional features. In addition, selection for multivalent binding functions may give rise to symmetry in globular, repeat, toroidal, and oligomeric proteins [142]. In this respect, it is noteworthy that the ricins, AbfB, and fascin β-trefoil families most prominently exhibit subdomain repetition, and these families are involved in the multivalent binding of ligands: carbohydrates in the case of the ricins [139] and AbfBs [156] and actin in the case of the fascins [157]. Ricin β-trefoils are involved in host-pathogen interactions, which often require multivalent binding achieved through symmetry [158]. Also, they exist as domains within rapidly evolving toxins [159], which may provide an increased opportunity to observe repetition events.

The repetition of subdomain modules to form internally symmetric structures may provide an inherent benefit for functional plasticity, as compared with oligomerization in a repeat protein. In an oligomeric protein, any mutations are necessarily present in all subunits of the oligomer, and this may limit the opportunity to acquire new or improved functions that require a combination of mutations, if any of the individual mutations along the way are functionally deleterious. By contrast, repetition of a subdomain module into a single larger protein with multiple initially identical functional sites means that at least one of these functional sites may be free to accumulate mutations that are initially deleterious to function but may eventually lead to novel or improved function [149], and this modified subdomain may then itself repeat to form a newly symmetric protein with amplified function.

Symmetric protein structures (due to internal repetition or oligomerization) are the rule rather than the exception in nature. The reasons for this have been the subject of much speculation [142, 151]. Both physical factors, as described above, and evolutionary mechanisms at the level of DNA replication may play important roles [142]. Using sequence analysis and rational design tools we have successfully reconstructed a fully internally symmetric protein sequence with the intended monomeric structure, which has the attractive features of being well folded, highly soluble, and functional, and exhibiting high thermal stability. The structure itself and the design strategy of combining bioinformatics and protein modelling should be useful for future studies of the origins and determinants of symmetric protein folds, as well as for designing proteins with desirable properties such as multivalent binding and high stability. In particular, the application of modular evolution or modular protein design may prove to be an elegant solution to incorporating functional properties into a scaffold while retaining sufficient stability, and recent work in addition to this study highlights the promise of such approaches [53, 127, 144, 160, 161]. Although there is still relatively little experimental data on the sequence repetition of subdomain modules in globular proteins, it seems likely that closer examination of the vast and ever-expanding protein sequence and structure databases will continue to provide further evidence for these processes in other symmetric folds.

2.6 Methods

2.6.1 Sequence dataset construction and analysis

All annotated β-trefoil domain sequences were retrieved from the National Center for Biotechnology Information (NCBI) using the Conserved Domain Database (CDD). All families annotated as β-trefoils by SCOP [162] and Pfam [124] with an available structure in the Protein Data Bank (http://www.pdb.org) [163] were included. See Table 2.1 for statistics on construction of the dataset. Sequences were parsed and their β-trefoil regions extracted according to the CDD information included in the NCBI's GenPept file. All β-trefoil domains in each protein chain were extracted, which resulted in an initial dataset of 5287 domain sequences. To remove redundancy, all domain sequences were grouped into clusters of highly similar sequences using the BLASTCLUST algorithm from the BLAST package [138] with default parameters (length coverage threshold = 0.9; score coverage threshold = 1.75). The longest sequence from each cluster was selected as a representative, and the remaining sequences were removed from the dataset. β-trefoil sequences were then parsed into their individual subdomain modules by aligning all sequences to their corresponding β-trefoil family HMM using the program HMMalign (http://hmmer.janelia.org), and dividing the sequences into three parts according to the repeat pattern evident within a representative structure. The representative structures used in subdomain module parsing are listed in Table 2.1. Sequences that were truncated and/or contained insufficient data were excluded by only including sequences containing three subdomain modules with lengths longer than 20 residues. The final dataset consisted of 3501 subdomain modules from 1167 β-trefoil domains. Subdomain modules were then clustered using a graph-based approach. First, an all-by-all BLAST search was performed, and any two repeats with E $<$ 1e-04 were connected. The results were visualized with the program Cytoscape (http://www.cytoscape.org). The choice of clustering parameter will change the evolutionary resolution of the analysis. A lower BLAST E-value will result in a larger number of clusters and require the detected subdomain-repetition events to be higher in similarity and thus more recent. Conversely, a higher E-value threshold will result in fewer clusters but identify potentially more ancestral subdomain-repetition events. The cutoff of 1e-04 was chosen as a reasonable middle-ground.

In addition to the 9 cases of subdomain repetition represented by the 9 clusters in Figure 2.3C, we looked for additional subdomain repetition events within each cluster. In order to identify these additional repetition events, we constructed a subset of domain sequences for which the internal subdomains aligned significantly to each other (E <1 e-04). This subset comprised 29 domains (87 subdomains), and included subdomains from the 9 representative sequences in Table 2.3 corresponding to the 9 clusters in Figure 2.3C. We then constructed a neighbour-joining tree from a multiple sequence alignment with gaps, and a Poisson correction, using Seaview [164] (Figure 2.9D). The tree was then used to organize the axes of a heatmap of pairwise similarity E-values to allow for the pairwise comparison of all subdomains in the tree.

2.6.2 Design of ThreeFoil

An overview of the design methodology is given in the Results section, with additional details below. The template sequence was divided into its three subdomain modules, each of length 47 amino acids, using the method employed for sequence dataset construction and analysis. The homologous sequences used in reconstruction were the 13 most closely related nonredundant sequences identified using BLAST [138]. These were split into their three constituent subdomain modules after multiple sequence alignment with the template using MUSCLE [133], giving 39 homologous subdomain modules (Figure 2.9). For Rosetta Design, an initial structure was needed. Three structures were generated through homology modelling using MODELLER [165] (http://www.salilab.org/modeller/) and the three most closely related structures, PDB IDs 1KNM [166], 2IHO [131], and 1YBI [167]. The sequence used to generate the homology models for step 3 of the reconstruction was from step 2 (Figure 2.4, Step 2), with the 16 gaps filled in by the most frequent residue based on the step 2 frequency calculation (see Results).

2.6.3 Expression and purification of ThreeFoil and OneFoil

The nucleotide sequence of ThreeFoil was synthesized and supplied in a pUC57 vector (GenScript). The sequence was subcloned into a modified pET-28a vector containing an N-terminal deca-histidine tag (modified from the original hexa-histidine tag). OneFoil was generated by annealing six overlapping oligonucleotides (Sigma Aldrich) followed by direct ligation of the oligos into linearized modified pET-28a. Both ThreeFoil and OneFoil were expressed in emphEscherichia coli after induction with IPTG (1 mM). Cells were harvested after 48 and 24 hr of growth at 37 °C and 25 °C for ThreeFoil and OneFoil, respectively. Both proteins were isolated as inclusion bodies and solubilized in buffered urea (6 M urea, 100 mM phosphate, and 10 mM tris, pH 8.1), bound to a Ni-NTA column, and eluted at pH 4.5. The purified protein was then refolded by dialysis (SpectraPor10) against 300 mM NaCl and 100 mM phosphate, pH 6.6 (the standard buffer used for all subsequent experiments) at a concentration of 0.15 mg/ml, and then concentrated to 12.5 and 1.0 mg/ml for ThreeFoil and OneFoil, respectively, using ultrafiltration (YM10 membranes, Amicon). Due to solubility limits, OneFoil could not be concentrated to the same levels as ThreeFoil. Molar extinction coefficients of 33,600 and 11,200 L mol^{-1} cm^{-1} for ThreeFoil and OneFoil, respectively, were determined using the method of Pace and co-workers [168] and used for determination of protein concentrations.

2.6.4 Structure determination and refinement

ThreeFoil was screened for crystallization conditions using the Index HT screen (Hampton Research). Crystals of ThreeFoil appeared after one month from sitting drops (2.4 M ammonium sulfate and 100 mM bis-tris, pH 6.5) at a protein concentration of 7 mg/ml and were soaked in the aforementioned solution with the addition of 25% (v/v) glycerol as a cryoprotectant before being flash-frozen in a -170 °C N_2 stream. Data were collected in-house at The Hospital for Sick Children in Toronto and processed using d*TREK [169]. The structure of ThreeFoil (deposited as PDB code 3PG0) was solved by Molecular Replacement and refined to 1.62 Å resolution using the Balbes molecular replacement server (http://www.ysbl.york.ac.uk/∼fei/balbes/) [170]. Balbes selected the C-terminal CBM13 domain (128 residues) of *Streptomyces lividans* xylanase 10A (PDB code 1KNM) as the best search model. All 8 possible space groups within the symmetry class were tried and the best molecular replacement solution with one monomer in asymmetric unit was obtained in space group P43212 with a certainty factor of 72% and R and R_{free} of 37.7 and 40.2, respectively, after initial rigid body and positional refinement. The quality of the electron density permitted most of the residues to be built. The structure was refined using the Phenix (http://www.phenix-online.org/) [171] software suite with manual rebuilding using Coot (http://www.biop.ox.ac.uk/coot/) [172]. Model validation was performed using Molprobity (http://molprobity.biochem.duke.edu/) [173] and Procheck (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) [174]. The refinement, validation and rebuilding steps were performed iteratively and were guided by the decrease in the value of R_{free}. The final model includes residues 2 to 141. Density of insufficient quality prevented the modelling of the N-terminal His-tag and Gly 1. The TLSMD server (http://skuld.bmsc.washington.edu/∼tlsmd/) was used to divide the protein into 9 TLS (Translation/Liberation/Screw) motion domains [175] and these were used in the final rounds of the Phenix refinement to determine anisotropic atomic displacement parameters. The data collection and refinement statistics are presented in Table 2.2.

2.6.5 Biophysical characterization

All samples of ThreeFoil and OneFoil were prepared in 300 mM NaCl and 100 mM phosphate, pH 6.6, and analyzed at ambient temperature unless otherwise noted. DLS measurements were made using a protein concentration of 12.5 mg/ml, with a 0.4 cm pathlength cuvette and a NanoSZ particle sizer (Malvern). SLS measurements were obtained at the same time as DLS for protein concentrations ranging from 0.7 to 12.5 mg/ml. SEC was performed using a Superose 12 HR 10/30 column (GE Healthcare), with a 0.5 ml/min flow-rate, with buffer supplemented with D-galactose (1.5 M). Fluorescence measurements were performed using a Flourolog322 (Spex) with excitation and emission wavelengths of 280 nm and 313 nm, respectively, and excitation and emission slit widths of 1 nm and 5 nm, respectively. DSC was performed at a ThreeFoil concentration of 0.6 mg/ml and a scan rate of 1 °C/min. One-dimensional ^{1}H-NMR spectra were acquired at 25 °C using a Bruker AVANCE 600 MHz spectrometer with a TSI probe and excitation sculpting for water suppression [176], with a ThreeFoil concentration of 12.5 mg/ml and a OneFoil concentration of 1.0 mg/ml. Glycan array analysis was performed as reported by the Consortium for Functional Glycomics [177] (http://www.functionalglycomics.org/fg/), using a ThreeFoil concentration of 0.2 mg/ml.

2.7 Supplemental Information

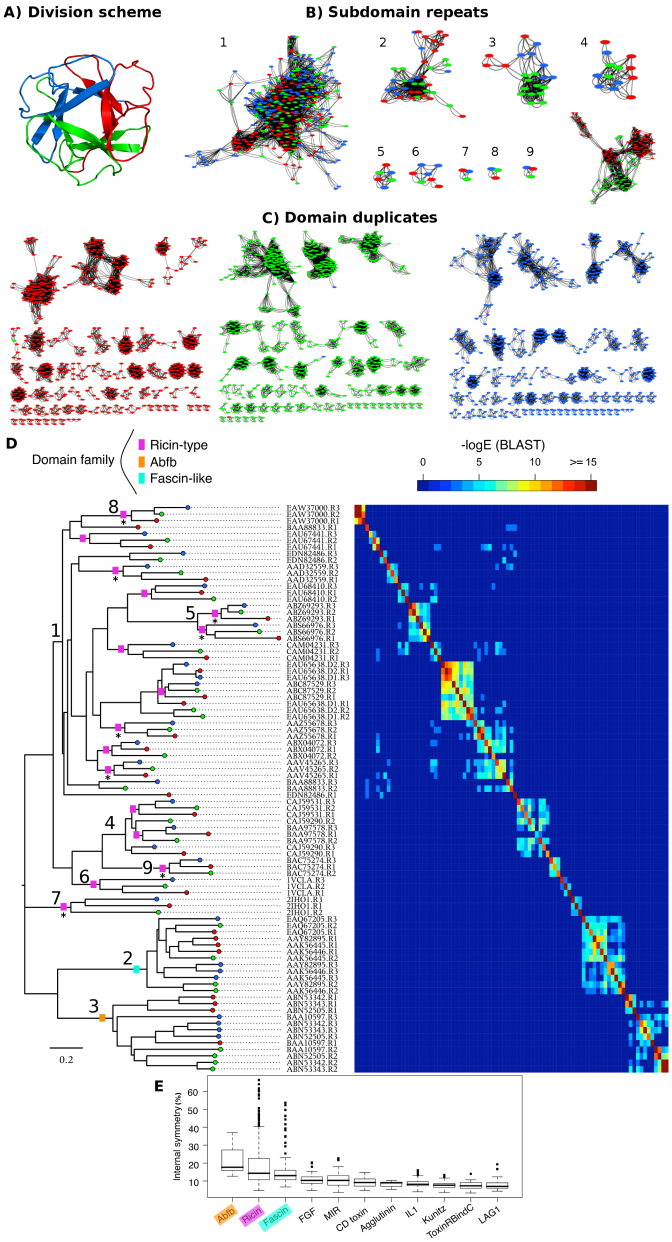
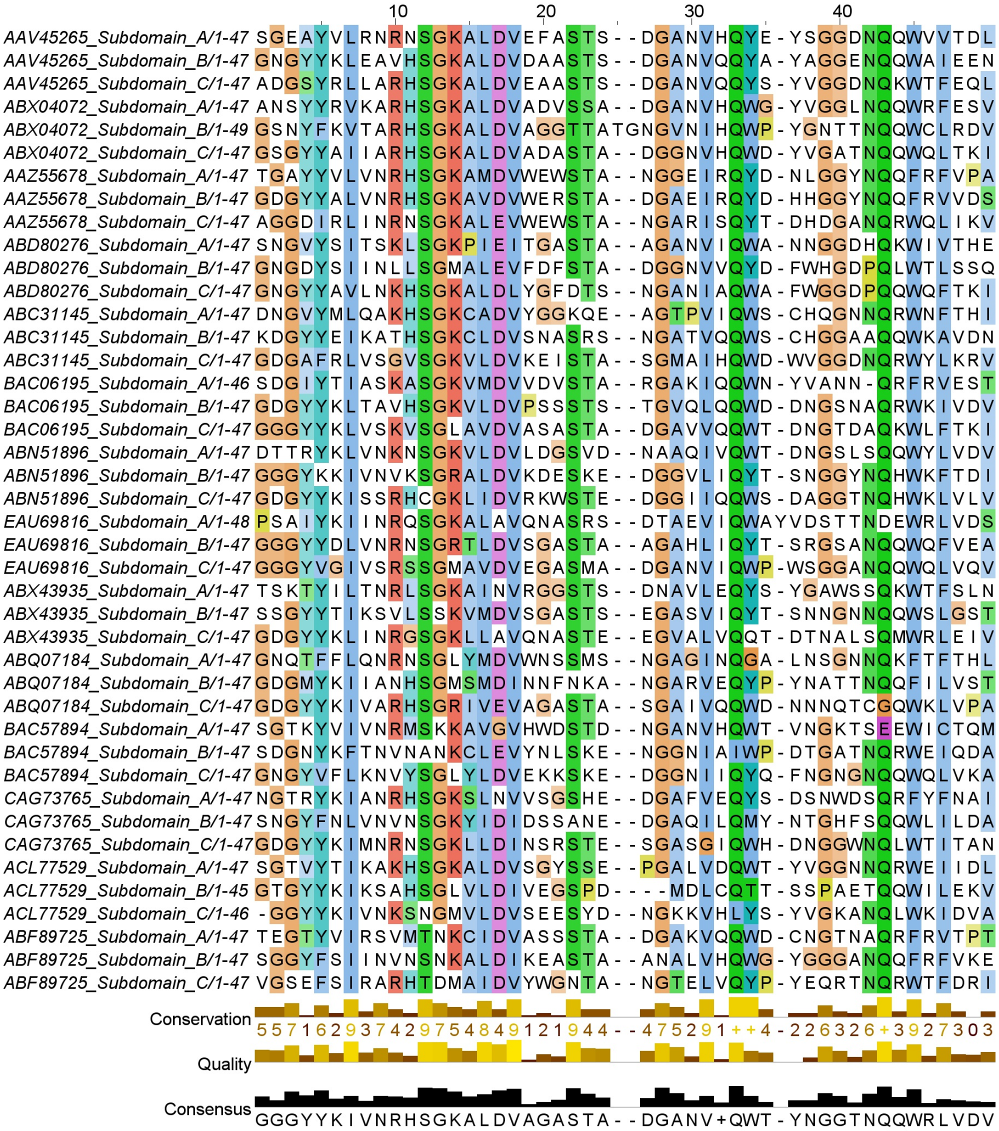


Figure 2.9: Sequence clustering and phylogenetic analysis of β-trefoil subdomain modules.

Individual subdomain modules represented by a sphere are colored as in Figure 2.1 and clustered according to sequence similarity as described in the Methods and Results. (A) Structural division of β-trefoil domains into three constituent symmetrical subdomain modules. (B) β-trefoil sequence clusters corresponding to evolution *via* subdomain repetition. Internal subdomains are more closely related to one another than to extant subdomains. (C) Clusters demonstrating whole domain duplication and divergence resulting in subdomains that are most closely related to the corresponding subdomain in homologous sequences. Cluster numbering corresponds to Table 2.3. (D) A neighbor-joining phylogenetic tree based on an alignment of the subdomains with the highest internal similarity (87 subdomains the 29 most internally symmetric domains), with pairwise similarity E-values visualized as a heat map. Clades corresponding to a particular cluster in panel B) are indicated by numbers in front of the clade branch-point. In the tree, subdomains are colored as per the above panels, and subdomain repetition events are indicated by boxes (colored corresponding to sequence family). This shows a total minimum of 18 subdomain repetition events. The subdomain sequences that have the highest internal sequence similarity are marked with an asterisk. (E) A box-plot of internal sequence similarity as determined by average % sequence identity between subdomains within a sequence for each β-trefoil family. There are more high-symmetry outliers in the ricin and fascin families compared to their overall symmetry distributions, indicating multiple sequence repetition events. Overall, the symmetry scores are high for the Abfb family but there are no high-symmetry outliers; this is consistent with a single ancestral subdomain repetition event followed by normal duplication and divergence. Family labels are colored as in panel D).



```
                                   10        20        30        40
AAV45265_Subdomain_A/1-47 SGEAYVLRNRNSGKALDVEFASTS--DGANVHQYE-YSGGDNQQWVVTDL
AAV45265_Subdomain_B/1-47 GNGYYKLEAVHSGKALDVDAASTS--DGANVQQYA-YAGGENQQWAIEEN
AAV45265_Subdomain_C/1-47 ADGSYRLLARHSGKALDVEAASTS--DGANVQQYS-YVGGDNQKWTFEQL
ABX04072_Subdomain_A/1-47 ANSYYRVKARHSGKALDVADVSSA--DGANVHQWG-YVGGLNQQWRFESV
ABX04072_Subdomain_B/1-49 GSNYFKVTARHSGKALDVAGGTTATGNGVNIHQWP-YGNTTNQQWCLRDV
ABX04072_Subdomain_C/1-47 GSGYYAIIARHSGKALDVADASTA--DGGNVHQWD-YVGATNQQWQLTKI
AAZ55678_Subdomain_A/1-47 TGAYYVLVNRHSGKAMDVWEWSTA--NGGEIRQYD-NLGGYNQQFRFVPA
AAZ55678_Subdomain_B/1-47 GDGYYALVNRHSGKAVDVWERSTA--DGAEIRQYD-HHGGYNQQFRVVDS
AAZ55678_Subdomain_C/1-47 AGGDIRLINRNSGKALEVWEWSTA--NGARISQYT-DHDGANQRWQLIKV
ABD80276_Subdomain_A/1-47 SNGVYSITSKLSGKPIEITGASTA--AGANVIQWA-NNGGDHQKWIVTHE
ABD80276_Subdomain_B/1-47 GNGDYSIINLLSGMALEVFDFSTA--DGGNVVQYD-FWHGDPQLWTLSSQ
ABD80276_Subdomain_C/1-47 GNGYYAVLNKHSGKALDLYGFDTS--NGANIAQWA-FWGGDPQQWQFTKI
ABC31145_Subdomain_A/1-47 DNGVYMLQAKHSGKCADVYGGKQE--AGTPVIQWS-CHQGNNQRWNFTHI
ABC31145_Subdomain_B/1-47 KDGYYEIKATHSGKCLDVSNASRS--NGATVQQWS-CHGGAAQQWKAVDN
ABC31145_Subdomain_C/1-47 GDGAFRLVSGVSGKVLDVKEISTA--SGMAIHQWD-WVGGDNQRWYLKRV
BAC06195_Subdomain_A/1-46 SDGIYTIASKASGKVMDVVDVSTA--RGAKIQQWN-YVANN-QRFRVEST
BAC06195_Subdomain_B/1-47 GDGYYKLTAVHSGKVLDVPSSSTS--TGVQLQQWD-DNGSNAQRWKIVDV
BAC06195_Subdomain_C/1-47 GGGYYKLVSKVSGLAVDVASASTA--DGAVVQQWT-DNGTDAQKWLFTKI
ABN51896_Subdomain_A/1-47 DTTRYKLVNKNSGKVLDVLDGSVD--NAAQIVQWT-DNGSLSQQWYLVDV
ABN51896_Subdomain_B/1-47 GGGYKKIVNVKSGRALDVKDESKE--DGGVLIQYT-SNGGYNQHWKFTDI
ABN51896_Subdomain_C/1-47 GDGYYKISSRHCGKLIDVRKWSTE--DGGIIQQWS-DAGGTNQHWKLVLV
EAU69816_Subdomain_A/1-48 PSAIYKIINRQSGKALAVQNASRS--DTAEVIQWAYVDSTTNDEWRLVDS
EAU69816_Subdomain_B/1-47 GGGYYDLVNRNSGRTLDVSGASTA--AGAHLIQYT-SRGSANQQWQFVEA
EAU69816_Subdomain_C/1-47 GGGYVGIVSRSSGMAVDVEGASMA--DGANVIQWP-WSGGANQQWQLVQV
ABX43935_Subdomain_A/1-47 TSKTYILTNRLSGKAINVRGGSTS--DNAVLEQYS-YGAWSSQKWTFSLN
ABX43935_Subdomain_B/1-47 SSGYYTIKSVLSSKVMDVSGASTS--EGASVIQYT-SNNGNNQQWSLGST
ABX43935_Subdomain_C/1-47 GDGYYKLINRGSGKLLAVQNASTE--EGVALVQQT-DTNALSQMWRLEIV
ABQ07184_Subdomain_A/1-47 GNQTFFLQNRNSGLYMDVWNSSMS--NGAGINQGA-LNSGNNQKFTFTHL
ABQ07184_Subdomain_B/1-47 GDGMYKIIANHSGMSMDINNFNKA--NGARVEQYP-YNATTNQQFILVST
ABQ07184_Subdomain_C/1-47 GDGYYKIVARHSGRIVEVAGASTA--SGAIVQQWD-NNNQTCGQWKLVPA
BAC57894_Subdomain_A/1-47 SGTKYVIVNRMSKKAVGVHWDSTD--NGANVHQWT-VNGKTSEEWICTQM
BAC57894_Subdomain_B/1-47 SDGNYKFTNVNANKCLEVYNLSKE--NGGNIAIWP-DTGATNQRWEIQDA
BAC57894_Subdomain_C/1-47 GNGYVFLKNVYSGLYLDVEKKSKE--DGGNIIQYQ-FNGNGNQQWQLVKA
CAG73765_Subdomain_A/1-47 NGTRYKIANRHSGKSLNVVSGSHE--DGAFVEQYS-DSNWDSQRFYFNAI
CAG73765_Subdomain_B/1-47 SNGYFNLVNVNSGKYIDIDSSANE--DGAQILQMY-NTGHFSQQWLILDA
CAG73765_Subdomain_C/1-47 GDGYYKIMNRNSGKLLDINSRSTE--SGASGIQWH-DNGGWNQLWTITAN
ACL77529_Subdomain_A/1-47 SGTVYTIKAKHSGKALDVSGYSSE--PGALVDQWT-YVGGNNQRWEIIDL
ACL77529_Subdomain_B/1-45 GTGYYKIKSAHSGLVLDIVEGSPD----MDLCQTT-SSPAETQQWILEKV
ACL77529_Subdomain_C/1-46 -GGYYKIVNKSNGMVLDVSEESYD--NGKKVHLYS-YVGKANQLWKIDVA
ABF89725_Subdomain_A/1-47 TEGTYVIRSVMTNKCIDVASSSTA--DGAKVQQWD-CNGTNAQRFRVTPT
ABF89725_Subdomain_B/1-47 SGGYFSIINVNSNKALDIKEASTA--ANALVHQWG-YGGGANQQFRFVKE
ABF89725_Subdomain_C/1-47 VGSEFSIRARHTDMAIDVYWGNTA--NGTELVQYP-YEQRTNQRWTFDRI
             Conservation 557162937429754849121944--475291++4-226326+3927303
                  Quality
                Consensus GGGYYKIVNRHSGKALDVAGASTA--DGANV+QWT-YNGGTNQQWRLVDV
```

Figure 2.10: Consensus design of ThreeFoil. A multiple sequence alignment of the template subdomains (first three sequences) with the subdomains of the 13 most closely related sequences (listed by NCBI accession number). The consensus was used to select amino acids at some of the non-symmetrical positions in the template.

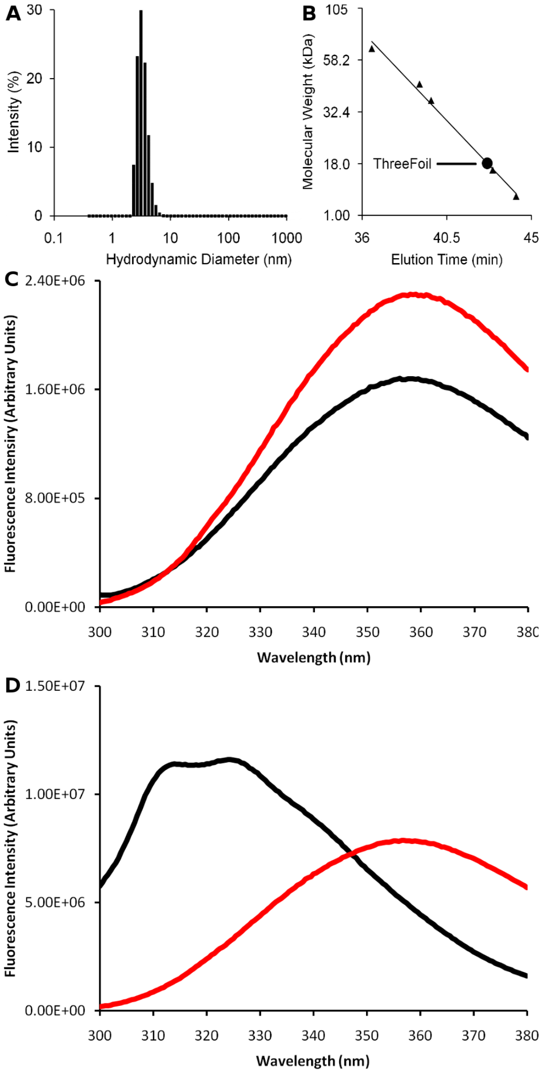


Figure 2.11: Additional biophysical characterization of ThreeFoil and OneFoil. Size measurements of ThreeFoil: (A) hydrodynamic diameter measured by DLS, and (B) calibrated size exclusion chromatography (SEC) (BSA, 66.3 kDa; ovalbumin, 44.3 kDa; b-lactoglobulin, 36.8 kDa; myoglobin, 16.7 kDa; cytochrome C, 12.4 kDa; ThreeFoil as black circle), consistent with expected size for a ThreeFoil monomer. (A) and (B) Fluorescence emission spectra shown for no denaturant (black line) and high denaturant (6 M guanidine hydrochloride, red line). (C) OneFoil shows the same emission maximum of ∼360 nm regardless of denaturant concentration, consistent with the fluorophores being solvent exposed in an unfolded peptide. (D) ThreeFoil demonstrates a very pronounced blue-shift between denaturant (peak at ∼360 nm) and no denaturant (peaks at 313 nm and 323 nm), consistent with a change in the environment of the tryptophan fluorophores from solvent exposed, to buried hydrophobic environment.

Table 2.3: Representative structures and alignments for subdomain repetition clusters

Representative sequence	Multiple alignment of repeats	Cluster	Family	# repeats in cluster	# (domains, repeats) for which R1, R2, R3 are present in cluster
gi|55229846|gb|AAV45265.1 *Haloarcula marismortui* 1008-1139	------YVLRNRNSGKALDVEFAST-SDGANVHQYEYS---------GGDNQQWVVTDLGNGYY -------KLEAVHSGKALDVDAAST-SDGANVQQYAYA---------GGENQQWAIEEN----- --ADGSYRLLARHSGKALDVEAAST-SDGANVQQYSYV---------GGDNQKW----------	1	Ricin-like	376	64, 192
gi|86165939|gb|EAQ67205.1 *Marinomonas sp.* 10-125	-------------NSKYVCAENAG-------KSALIAN---------RERVGAWETFKVIPL-- --KGNKFALQA-CNGKYVCAERKG-------ANSLIAN---------RDKVGAWETFEWVN--- --KGNRKGFKAGCNGKHVCAEGGG-------AKALIAN---------RDNFDVWESFDV-----	2	Fascin-like	40	10,30
gi|125714851|gb|ABN53343.1 *Clostridium thermocellum* 479-613	---------------PAVGLQSYNYPNRYVRHADFDAR---IDENVTPLEDSQWRLVPGLA--- -----------NSSEGYVSIQSVNYPGYYLRHWDYDFRLDKNDGTTIFAEDATFKLVPGLAD-- --------------PSCVSFQSYNYPDRYIRHYGYLLKLERI-STDLDRQDATFLI--------	3	AbfB	21	4, 12
gi|111147635|emb|CAJ59290.1 *Frankia alni* 19-127	-------------TSKCLDSNGKGA------VYALGCN---------GGPYQSWVSSQLNFGD- -------QIQDVRTGRCLDSNGAKR------VYTLPCN---------GGSYQQWQVTDQ----- --GPFGYQIQNVATGFCLDSNGGGS------VYTHVCN---------GGNYQRW----------	4	Ricin-like	13	2, 6
gi|154159760|gb|ABS66976.1 *Xanthobacter autotrophicus* 36-168	----GRVQIANANSDLCLSPAGGTG-NQNEQTVQYHCD---------THPSRAWVIEPVEGNIV -------KIRNVNSNLCLTVAGGNS-DRNTPSVQYSCD---------DHPSRRWLYAPF----- --DGGLFRLVNVNSGLCLTIAGGST-GLNQTAVQFPCD---------EHPSRF-----------	5	Ricin-like	6	2, 6
gi|55670423|pdb|1VCL|A *Cucumaria echinata* 153-283	---LFYGRLRNEKSDLCLDVEGSDGKGN---VLMYSCE---------DNLDQ-WFR--YYE--- -----NGEIVNAKSGMCLDVEGSDGSGN---VGIYRCD---------DLRDQMWSRPNAYC--- --NGDYCSFLNKESNKCLDVSGDQGTGD---VGTWQCD---------GLPDQRFKWVF------	6	Ricin-like	7	1, 3
gi|149242174|pdb|2IHO|A *Marasmius oreades* 2-156	SLRRGIYHIENAGVPSAIDLKDGSS-SDGTPIVGWQFT------PDTINWHQLWLAEPIPN--- --VADTFTLCNLFSGTYMDLYNGSS-EAGTAVNGWQGT------AFTTNPHQLWTIKKSSD--- ---GTSYKIQNYGSKTFVDLVNGDS-SDGAKIAGWTGT------WDEGNPHQKWYFNRM-----	7	Ricin-like	3	1, 3
gi|119455865|gb|EAW37000.1 *Lyngbya sp.* 45-191	-------------SGKCIDVSGAPGRATGSKLQLWDCELSGLNPDNNSPSDQRWMITNDGF--- --------IKNTLSGKCIDVAGAPGTENGSPLQLWDCELSGRNRDNGSPTDQIWTITS------ -----DGFIMNRLSQKCIDVAGAPGQENSSALLLWDCELSGRNQDNGSSTDQIW----------	8	Ricin-like	3	1, 3
gi|29611230|dbj|BAC75274.1 *Streptomyces avermitilis* 120-218	-------------TLKCLDGSSRG-------IRLLKCN---------DSKYQRWQASPEDY--- -------SFRNVATLTCLDGSSRG-------LRLVKCN---------GSRYQAWVWRG------ ------DELHNAVLGTCLDGSSRG-------VRLVKCN---------DSKYQHW----------	9	Ricin-like	3	1, 3

Chapter 3

Subdomain Repetition Provides an Evolutionary Advantage

3.1 Context

The work in Chapter 2 began from a desire to test the evolutionary hypothesis that many modern domains originated from the repetition of smaller subdomain modules. Surprisingly, we discovered that in addition to being a feasible ancient event, the process of domain formation from subdomain-repetition is ongoing, at least for the β-trefoil fold. In the published work [44], we hypothesized that ongoing domain formation *via* subdomain repetition could allow for rapid evolution of new multivalent binding function, since repetition events were observed most commonly in families associated with host-pathogen interactions. The idea was left largely as an untested hypothesis at the time.

In parallel, Kyle Trainor and I repeated the bioinformatics analysis from Chapter 2 on a different fold, the β-prism. We found that, like the β-trefoil, there was evidence for ongoing domain-forming repetition events. Given the apparent generality of this mechanism in folds involved in multivalent binding. I wanted to quantify our previous arguments that this should allow for rapid evolution of multivalent binding, particularly within a host-pathogen context. To accomplish this I used genetic algorithms, which function by mimicking the genetic mechanisms of evolution. This chapter is a brisk description of that work.

3.2 Summary

The type I β-prism fold functions in host-pathogen interactions involving carbohydrate binding and possesses internal structural symmetry, suggesting emergence from a subdomain amplification event. Evidence from sequence analysis and clustering is presented which suggests that in addition to the ancient fold-forming event, at least one recent subdomain amplification event has occurred to regenerate a highly symmetric β-prism structure. This kind of ongoing evolution by subdomain amplification has previously been noted in both the β-propeller and β-trefoil folds, which also prevalently function in host-pathogen interactions through multivalent binding. I use genetic algorithms to demonstrate that subdomain amplification provides a considerable advantage in evolving multivalent binding to symmetric ligands like cell-surface glycans. Overall, the capacity of the β-prism fold to emerge *via* subdomain amplification, combined with a structure that places three symmetric variable binding loops in close proximity to one another, suggests this fold may be an attractive design candidate for diverse binding functions.

3.3 Introduction

As many as 20% of all protein domains possess internal structural symmetry [43]. These domains may have emerged from an ancient founding event in which a smaller subdomain module was amplified through successive duplication and fusion events [178, 122]. The size of these subdomain modules in modern proteins is typically ∼30-40 amino acids, agreeing well with the predicted size of ancient RNA-binding proteins which are thought to have been the early precursors to modern proteins [40, 179]. For some time it had been the general consensus that amplification of subdomain modules to form modern globular proteins was limited to being an ancient event. Evidence gained from large-scale analysis of β-propeller sequences, however, revealed that the sequence identity between the modules or "blades" of some modern propellers was extremely high, suggesting that fold formation from modular amplification is an ongoing process [136]. β-propeller blades — composed of a 4-stranded meander — may be particularly amenable to such amplification and could have formed other closely related folds [180]. As seen in Chapter 2, the β-trefoils also show evidence of ongoing domain formation by amplification of a 4-stranded module [44].

β-propellers and β-trefoils frequently function as multivalent glycan binders involved in host-pathogen interactions [44, 161], though β-propellers cover a very broad range of binding and enzymatic functions [181]. Both β-propellers and β-trefoils have been designed *de novo* to have perfect sequence symmetry, demonstrating the evolutionary feasibility of an emergence by amplification scenario [127, 44, 45]. Interestingly, there exists another fold heavily involved in glycan binding for host-pathogen interactions which may have originated from the ancient amplification of a 4-stranded element, the Type I β-prism (henceforth referred to simply as β-prism) [182, 183] (Figure 3.1). β-prisms may be best known as commercially relevant *Bacillus thuringiensis* toxins [184], a class of highly versatile insecticidal toxins which are used in genetically modified crops. Another large family of β-prisms come from seed lectins and other plant lectins which have been shown to posses potent antiviral activities [185, 186]. Both the insecticidal and antiviral properties of β-prisms appear to originate from a mannose-binding specificity [184, 185, 186], though galactose binding specificity has been observed in some cases [187].

Using the same bioinformatics analysis presented in Chapter 2, evidence for the ongoing evolution of β-prisms by subdomain module amplification is shown. In Chapter 2 it was suggested that such ongoing evolution could be beneficial within the context of an evolutionary arms race, as it might allow for rapid evolution towards multivalent targets such as cell surface glycans [44]. Using genetic algorithms to mimic the functional evolution of β-prisms to hypothetical glycan targets, I show that indeed amplification events, even at a relatively low rate, can dramatically reduce the number of generations needed to obtain high fitness to a new glycan target. Therefore, high sequence symmetry could be used to highlight genes involved in relatively new host-pathogen interactions. Furthermore, modelling results, such as optimum rates, might be useful in directed evolution of proteins against multivalent targets such as glycosylated viral receptors [185, 186, 188] or the gastrointestinal cells of agricultural pests [184].

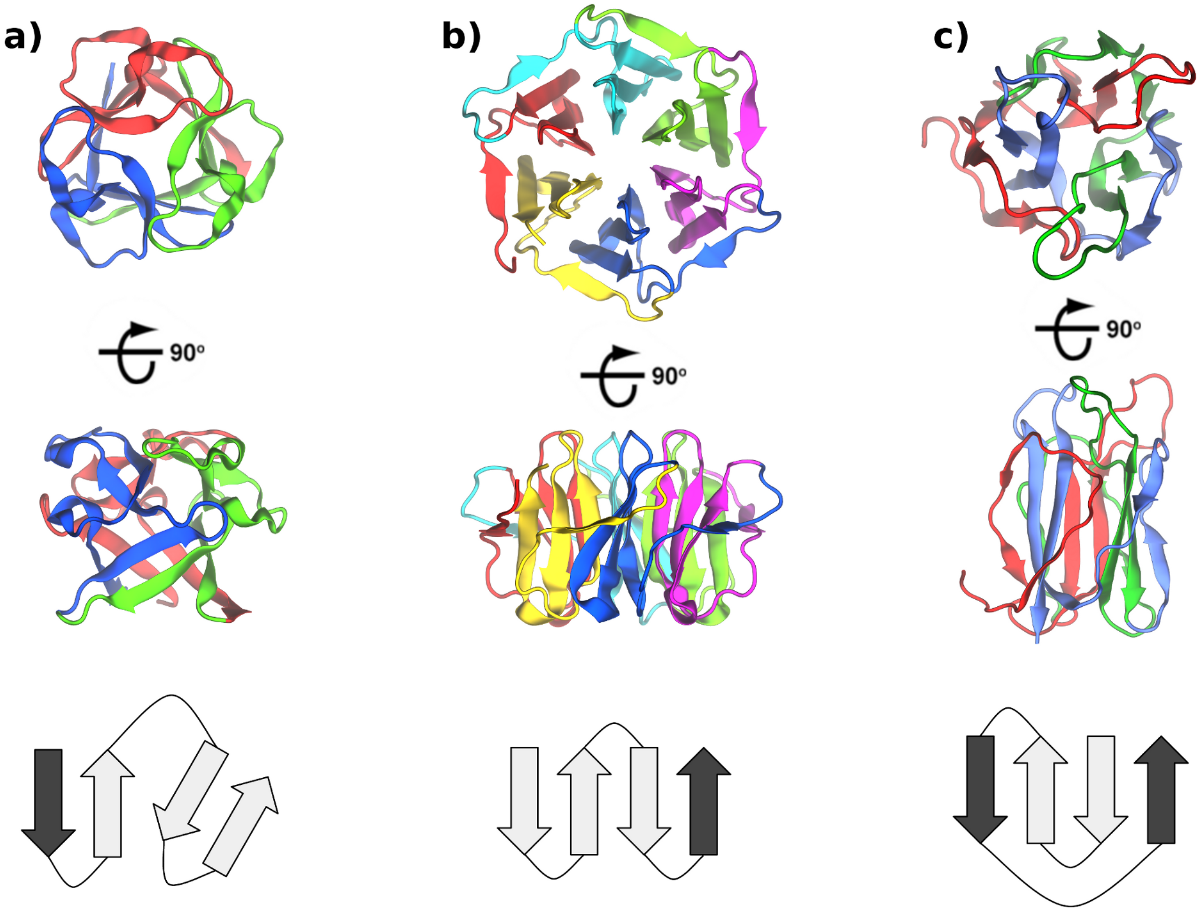


Figure 3.1: Folds evolving through ongoing subdomain amplification. The β-trefoil **(a)** and β-propeller **(b)** have previously been shown to evolve through ongoing subdomain amplification. In this work the same is shown for the β-prism **(c)**. A view along the axis of structural symmetry is shown for each fold (top), and perpendicular to the axis (middle). Repeating sequence modules are each colored differently. The repeating structural module is shown as a secondary structural diagram (bottom). Strands in the structural module are shaded to indicate they originate from different sequence modules. In all cases there is some degree of circular permutation between the sequence and structural repeats, which may serve to enhance folding cooperativity and kinetic stability by making the fold more topologically complex.

3.4 Results

3.4.1 Ongoing subdomain repetition in the β-prism fold

By clustering β-prism subdomain sequences, we show that while the majority are related by whole domain duplication and divergence, a single family of sequences homologous to Vitelline Outer Membrane protein I (VMO-I) have recently emerged *via* subdomain amplification (Figure 3.2).

Sequences known to form β-prisms were used to find additional homologs and guide their decomposition into subdomain modules (see Methods) (Figure 3.2a). Upon clustering subdomains by sequence similarity, two patterns are possible. In the case of whole domain duplication and divergence, which is generally accepted as the common mode of gene evolution [135], two homologous whole domains will show similarity between corresponding subdomains, which will cluster together. Non-corresponding subdomains, will not be similar and not cluster together (Figure 3.2b, case on the right). This occurs because, in the time since the ancient fold forming amplification event, neutral drift will have abolished much of the identity between subdomains. For two whole domains related by a comparably recent duplication and divergence event, there will exist considerable identity between corresponding sequence positions, and therefore, corresponding subdomains will be highly similar. On the other hand, if a new β-prism family has formed from a recent subdomain amplification event, the resulting whole domains will have detectable identity between non-corresponding subdomains, which will cluster together. In fact, if the family has been populated by subsequent duplication and divergence events, then all subdomains both corresponding and non-corresponding may cluster together (Figure 3.2b, case on the left), since all subdomain sequences still resemble the family's original founder subdomain.

Whole domain duplication and divergence is responsible for the majority of extant β-prism sequences (Figure 3.2c, clusters of a single color), where the majority of subdomain sequences cluster with the corresponding subdomain from related β-prisms. Nevertheless, there exists a single cluster/family in which all three subdomains cluster together (Figure 3.2c, leftmost cluster of all three colors). This cluster is represented by a single known structure, VMO-I (PDB: 1VMO), isolated from the outer vitelline membrane of hen's eggs. Localization in the outer membrane and co-localization with lysozyme suggests a role in pathogen defence [182]. Interestingly, while the exact function of VMO-I is not known, close homologs in mammals are found in tears and other secretions, strengthening the suggestion of a role as a host-pathogen defence protein [189].

3.4.2 Subdomain repetition speeds evolution of multivalent ligand binding

We had previously hypothesized that recent amplification events leading to proteins with a capacity for multivalent binding, are driven by competition within the context of an evolutionary arms race between host and pathogen. This was based on the observation that β-trefoil families with the highest preponderance of recent amplification events were those involved in host-pathogen interactions [44]. Additionally, the most symmetric natural 5-bladed propeller with known structure, Tachylectin-2 (PDB: 1TL2) functions as an innate host immunity protein in horseshoe crabs [152]. That VMO-I is most likely involved in some form of pathogen defence [182, 189] lends additional weight to this hypothesis.

To provide a more quantitative demonstration that domain evolution by subdomain amplification leads to rapid evolution of multivalent binding function, I modelled this using genetic algorithms [190]. Starting with a pool of sequences (of 2-6 subdomain repeats) having "weak" binding towards either a symmetric or asymmetric multivalent ligand (with matching valency), the interplay of recombination/crossover, and subdomain amplification rates on the number of generations required to obtain "strong" binding was tested (see methods). Increasing the rate of recombination/crossover led to more rapid evolution towards both symmetric and asymmetric ligands, with a rate of 1.0 (always mating with a crossover event) leading to a 33% reduction in the number of generations needed to acquire strong binding (Figure 3.3a,b). Given the prevalence of sexual mating, such an improvement is to be expected. Amplification from a subdomain module, however, had a dramatically different effect depending on the symmetry of the ligand. In the case of an asymmetric ligand, a low rate of subdomain amplification (∼0.05 to 0.10) led to a marginal, yet reproducible, reduction (∼5%) in the number of generations needed to obtain strong binding. Higher rates of subdomain amplification, in the context of an asymmetric ligand, led to much slower evolution with twice as many generations needed at a subdomain amplification rate of 0.85 (Figure 3.3d). When evolving binding to a symmetric ligand, however, subdomain amplification led to the most rapid evolution observed. The reduction in the number of generations followed an exponential decay, with a low amplification rate of 0.10 already leading to a 33% reduction (the same as a recombination rate of 1.0) and higher amplification rates resulting in upwards of 50% reduction (Figure 3.3c).

Mixing recombination and subdomain amplification, such that both could be used throughout the generations but were mutually exclusive when generating any particular offspring, showed the optimum combination of rates for each particular mechanism (Figure 3.3e,f). Specifically, for an asymmetric ligand having a mixture of a 0.9 rate of recombination with a 0.1 rate of subdomain amplification provided the best results, though pure recombination at a rate of 1.0 was nearly as good. For a symmetric ligand a 0.25 rate of recombination and 0.75 rate of subdomain amplification provided a nearly 66% reduction in the number of generations needed to see strong binding emerge.

The above trends held regardless of the valency of the protein and ligand, though the improvement gained from subdomain amplification was more striking at higher valencies (e.g. when simulating a 6-bladed β-propeller binding a glycan with 6-branches as compared with a β-prism binding a glycan with 3-branches) (Figure 3.4). These results may prove valuable in searching for proteins involved in host-pathogen interactions, where such proteins may be detected from their abnormally high level of internal sequence symmetry. This could be particularly valuable when combined with metaproteomics, to find novel microbicidal toxins for instance. Furthermore, the quantification provided here for optimal rates of recombination and subdomain amplification may be useful in generating successful directed evolution strategies.

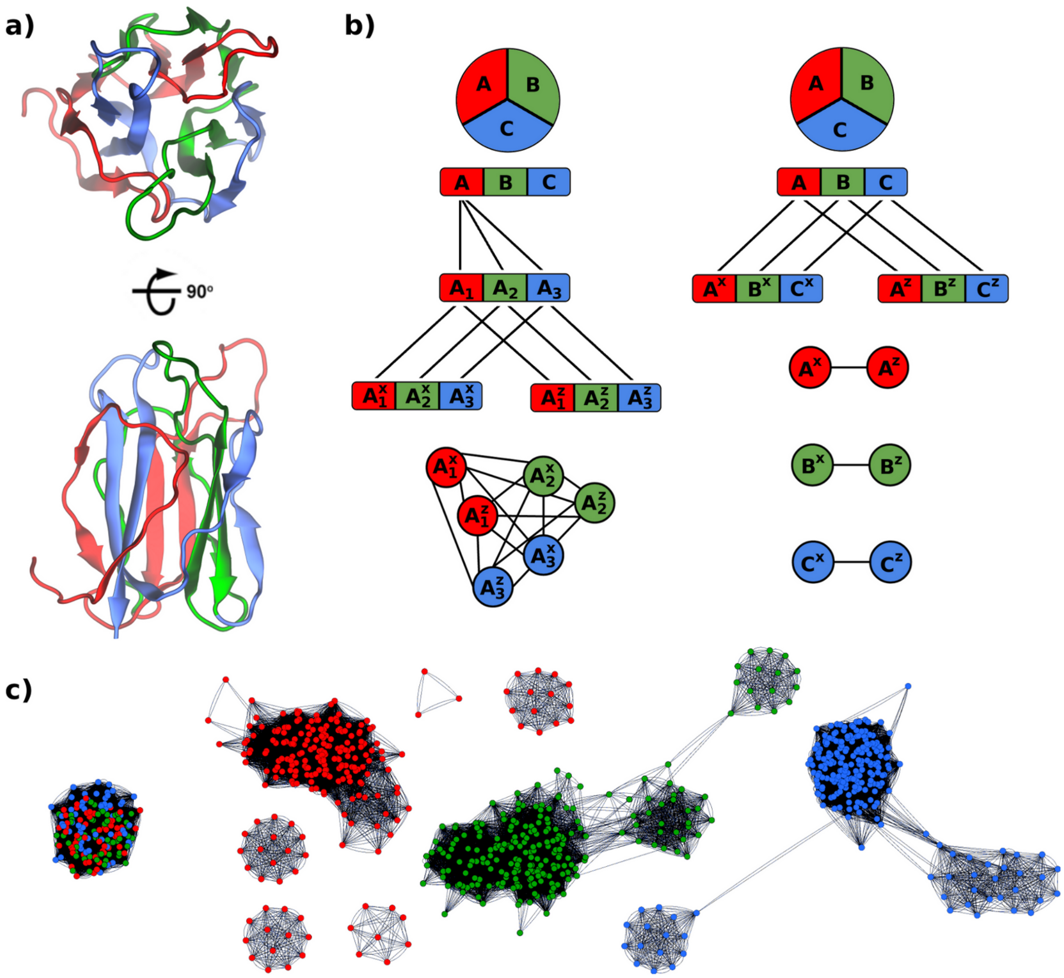


Figure 3.2: The β-Prism fold shows evidence of ongoing subdomain amplification events. **(a)** Breakdown of the β-prism structure into the three consecutive subdomain modules (red, green, blue), shown along the axis of symmetry (top), and perpendicular to it (bottom). **(b)** Two evolutionary scenarios for extant β-prism evolution are outlined. In both scenarios, the subdomain modules of the progenitor sequence do not share detectable internal homology (owing to neutral drift since the ancient fold-forming event). In the case of simple duplication and divergence (right), subdomain modules of the descendants must also lack detectable internal homology (e.g. A^x and B^x), but homology between corresponding modules (e.g. A^x and A^z) will be readily detectable if the duplication and divergence event was recent, and these subdomain modules will cluster together. In the case of subdomain amplification (left) the first descendant, which is formed from a new amplification event now shares complete homology between all three of its subdomains (A_1, A_2, A_3). Thus, after a subsequent duplication and divergence event occurs, the descendants will have detectable homology between all subdomain modules, which will cluster together. **(c)** The network layout after clustering all β-prism subdomain module sequences shows that the majority have evolved from duplication and divergence of an early member of the fold, but there has been at least one recent subdomain amplification event generated the sequences which form a unique cluster with high internal symmetry (leftmost cluster of all 3 colors).

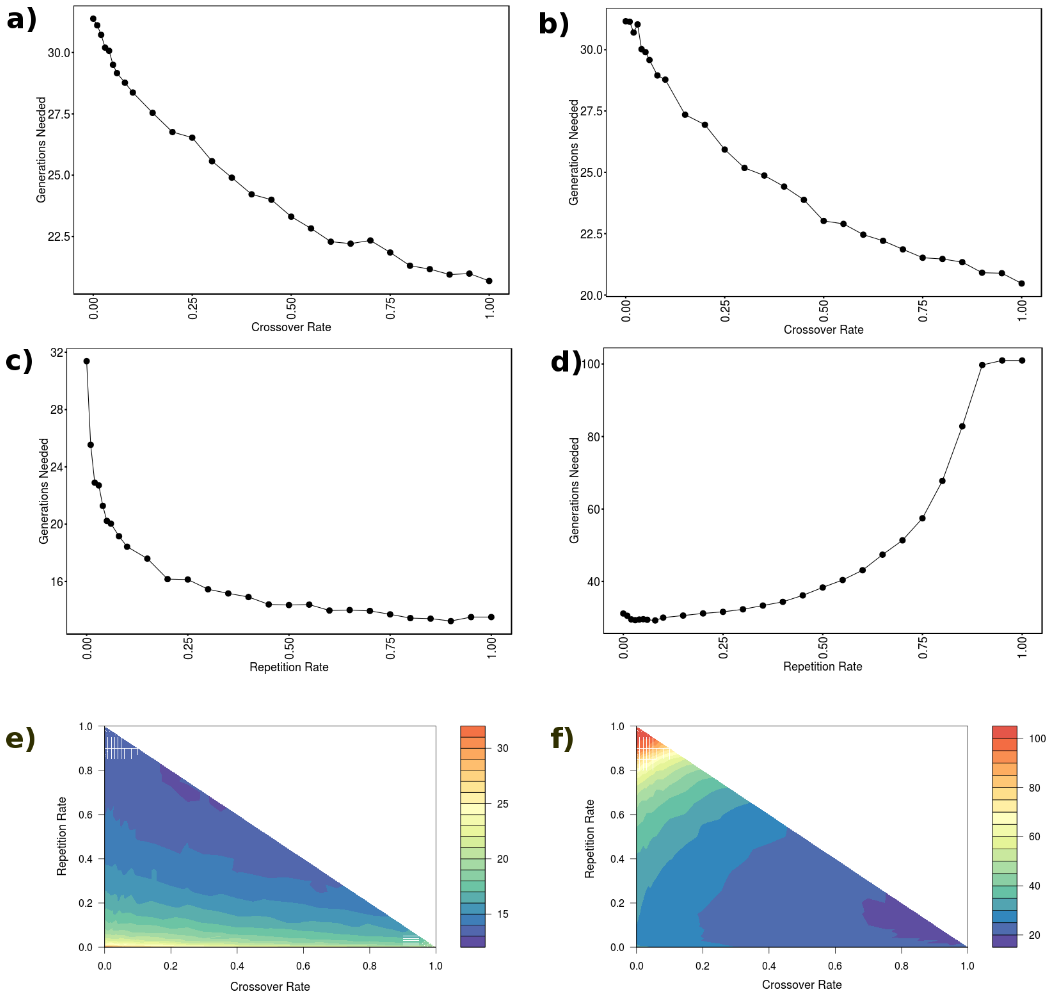


Figure 3.3: Subdomain amplification allows for rapid evolution of multivalent binding specificity to symmetric ligands. The number of generations required for the genetic algorithm to move from weak binding (fitness of 0.1) to strong binding (fitness of 0.9) is shown for different ligand symmetries and rates of recombination (crossover) and subdomain amplification. Recombination is equally beneficial whether the ligand is symmetric **(a)** or asymmetric **(b)**. By contrast, subdomain amplification is detrimental (beyond very low rates) when attempting to evolve binding to an asymmetric ligand **(d)**, but extremely beneficial even at moderate rates when evolving binding to a symmetric ligand **(c)**. When both mechanisms are combined (but mutually exclusive when generating any particular offspring sequence), it can be seen that for a symmetric ligand the optimal combination involves mostly subdomain amplification with a limited chance of recombination **(e)** and *vise versa* for an asymmetric ligand **(f)**.

3.5 Discussion

As the β-prism fold has emerged from amplification of a single module more than once, and a high degree of sequence symmetry is still evident in extant sequences, it should be possible to design a β-prism binding scaffold using symmetric design. Such an approach can dramatically reduce the size of the design space that needs to be searched and lead to a high likelihood of success, as seen for both the β-trefoils and β-propellers [1].

The β-prism fold has several structural features which make it an attractive design target. First, the overall topology is very complex, which is known to be correlated with slow unfolding and kinetic stability. For instance, the long-range order of VMO-I is 7.7, higher than any single domain proteins known to be kinetically stable, and notably high given the range for globular proteins of 0 to 8 [70]. As such, it may be an excellent scaffold when resistance to: proteases, detergents, high temperature, and other harsh conditions is needed [70]. Localization of β-prisms to extracellular environments [182, 189, 184], and the fact that digestive enzymes activate, rather than destroy, *Bacillus thuringiensis* toxins using the β-prism as a targeting domain [184], strongly suggest the aforementioned resistances are already being taken advantage of in natural proteins. Second, while β-trefoils and β-propellers offer alternative multivalent binding scaffolds, the β-prism is unique in the close spatial distribution of its three putative binding sites, which all occur at one end of a protein which has an oblong shape (Figure 3.1). Notably, this structure means that the variable binding loops made by each subdomain module are in close proximity to one another, yielding the possibility of cooperative binding as seen for classically adaptable binding scaffolds like the immunoglobulins [184]. Therefore, the β-prism represents an exciting scaffold for the potential design of both multivalent and monovalent binding proteins.

In our earlier work [44], we noted that β-trefoil families involved in recent subdomain amplification events were disproportionately involved in host-pathogen interactions. The work here quantitatively demonstrates why this is the case. Namely, subdomain amplification allows for dramatically faster evolution of binding towards multivalent targets. In particular, glycan structures which cover the surface of viruses, archea, bacteria, and eukaryotes are either themselves multivalent or present in large enough numbers that multivalent binding is an effective interaction strategy [191]. Thus, recent amplification events, which can be detected from sequence data, may be useful for finding proteins involved in host-pathogen interactions, such as novel anti-virals [185, 186, 188]. Alternatively, symmetric scaffolds such as the β-prisms, β-trefoils, or β-propellers might be more effectively engineered towards multivalent targets by taking advantage of the benefits of subdomain amplification.

3.6 Methods

3.6.1 Subdomain assignment and clustering

An set of 29 unique β-prism structures was obtained from the PDB (http://www.rcsb.org/) using the structural similarity search and then eliminating all structures of the same protein. Sequences for these structures were then used in a BLAST [138] search with a 10^{-4} E-value cutoff to find and align homologous sequences. This yielded a total of 606 sequences for analysis. Each structure was manually divided into its subdomains by cutting in the middle of the loop between the 4th and 5th and 8th and 9th strands. The resulting cut-points were used to divide the aligned homologous sequences, yielding the full set of subdomain sequences. The subdomain sequences were then clustered based on an E-value cutoff of 10^{-4} and a graphical representation produced using a force-based layout in Cytoscape (http://www.cytoscape.org/). In this representation, any two subdomain sequences homologous within the E-value cutoff, are connected by a spring with a particular force constant that scales with the level of homology. All connected subdomains are then clustered by applying a general repulsive force between all subdomains, while the force connecting related subdomains acts to counter this. The system is then integrated over time under an equilibrium position is established.

3.6.2 Genetic algorithm setup and execution

Each protein sequence was constructed to have 2-6 binding sites (depending on the valency being simulated). Each binding site contained 10 interactions used to determine the binding fitness, for which there were six possible properties: hydrophobic, aromatic, polar-negative, polar-positive, charged-negative, and charged-positive. This was chosen to roughly mimic a medium-sized binding site for which 10 amino acids would collectively determine binding specificity. Multivalent ligands had between 2-6 antennae (depending on the valency being simulated) and the same six possible properties as the binding sites. Fitness was determined simply by the fraction of binding site properties that matched with the corresponding ligand properties (i.e. the property of the third "amino acid" within a given binding site would need to match with the third property within the corresponding ligand antenna). This was designed to recapitulate in a rough manner the physical properties of multivalent ligand binding. One possible complication is that in the case of an asymmetric ligand I did not allow non-corresponding binding sites and ligand antenna to be paired, which of course would be possible in a natural physical setting where the best matches would pair by virtue of free energy. Nevertheless, such an addition would simply improve the overall evolutionary rate of the asymmetric ligand during recombination, and would not impact the results seen with subdomain amplification. Thus the comparative benefit of subdomain amplification for symmetric over asymmetric ligands would remain unaffected.

3.7 Supplemental Information

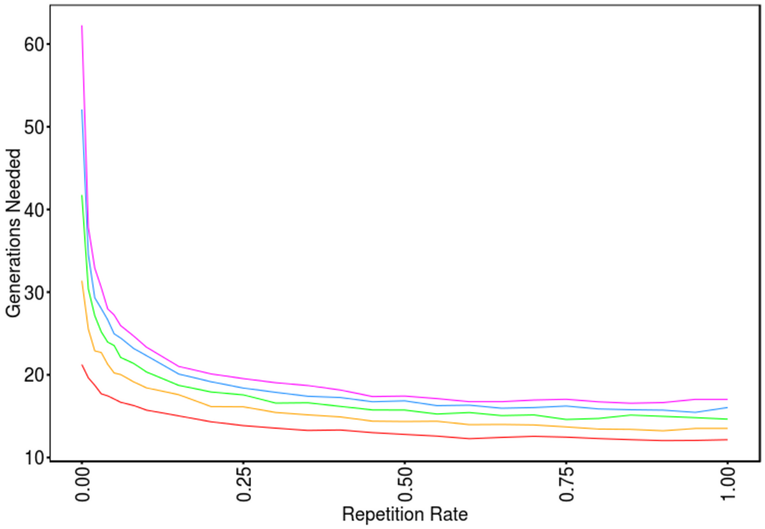


Figure 3.4: Subdomain amplification yields increased benefits with higher valency structures. The number of generations required for the genetic algorithm to move from weak binding (fitness of 0.1) to strong binding (fitness of 0.9) towards a symmetric ligand depending on the valency of the ligand and protein is shown. The benefit derived from repetition is greatest at higher valencies like 4 (green), 5 (blue), and 6 (magenta), as seen in β-propellers. The benefit is intermediate for a valency of 3 (orange) as seen for β-trefoils and β-prisms, while the benefit is much lower for a valency of 2 (red).

Chapter 4

Topological Complexity and Protein Kinetics

4.1 Context

This chapter is a pre-print version of “Protein unfolding rates correlate as strongly as folding rates with native structure”, published in Protein Science in 2015, of which I am the first author [78]. The key finding of this work is that the topological complexity of a protein’s structure, which can be quantified by simple measures like Contact Order, not only predicts folding rates, but also unfolding rates. It had been known for some time that structures with a high topological complexity folded more slowly than their structurally simpler counter-parts, but existing work looking at unfolding rates had concluded that there was no correlation. I found that, in fact, the correlation is just as strong between the two, but owing to the larger variance in protein unfolding rates compared with folding rates, previous studies which used much smaller datasets had not been able to detect this correlation. Interestingly, the correlation found here, when using a dataset of >100 proteins, is quite strong (R ∼0.8). This suggests protein design/engineering avenues for those seeking particularly fast or slow unfolding rates. In particular, slow unfolding rates may provide benefits to survival under harsh conditions, and this paper laid the groundwork for a later paper presented in the next chapter of this thesis.

I performed all data collection and analysis. The data was interpreted, and the manuscript written, by: me, Shachi Gosavi, and Elizabeth M. Meiering.

4.2 Summary

Although the folding rates of proteins have been studied extensively, both experimentally and theoretically, and many native state topological parameters have been proposed to correlate with or predict these rates, unfolding rates have received much less attention. Moreover, unfolding rates have generally been thought either to not relate to native topology in the same manner as folding rates, perhaps depending on different topological parameters, or to be more difficult to predict. Using a dataset of 108 proteins including two-state and multistate folders, we find that both unfolding and folding rates correlate strongly, and comparably well, with well-established measures of native topology, the absolute contact order and the long range order, with correlation coefficient values of 0.75 or higher. In addition, compared to folding rates, the absolute values of unfolding rates vary more strongly with native topology, have a larger range of values, and correlate better with thermodynamic stability. Similar trends are observed for subsets of different protein structural classes. Taken together, these results suggest that choosing a scaffold for protein engineering may require a compromise between a simple topology that will fold sufficiently quickly but also unfold quickly, and a complex topology that will unfold slowly and hence have kinetic stability, but fold slowly. These observations, together with the established role of kinetic stability in determining resistance to thermal and chemical denaturation as well as proteases, have important implications for understanding fundamental aspects of protein unfolding and folding and for protein engineering and design.

4.3 Introduction

Extensive experimental [192, 193] and theoretical [194, 195, 196] research has been conducted to understand protein folding rates. In seminal studies, Plaxco *et al.* found that a simple measure of the topology of the native state, the Relative Contact Order (RCO), correlated well with folding rates for a small set of monomeric proteins that showed two-state behaviour [197]. Later, they revised their conclusions to show that Absolute Contact Order (ACO), which in addition to topology includes effects of protein length, correlated better for a larger dataset including multistate folders [198]. It has long been noted that folding rates depend on protein length, but the quantitative, physical basis for this dependence remains under investigation [199, 200, 201, 202, 203, 204] and additional studies have demonstrated that consideration of both topology and length leads to improved correlations compared to topology alone [205, 206]. Here we use structural complexity as a broad term to encompass the complexity imparted by both the topology and the length of the protein, and rate, when referring to the unfolding and folding rate constants (k_u and k_f respectively).

The correlation of ACO with folding rates suggests that while the transition state for folding lacks much of the well defined structure of the native state, it nevertheless has a broadly similar structure and complexity, as has been suggested from theory and simulation [196, 206, 207, 208] and demonstrated experimentally [209]. Many alternative measures of native state structural complexity have also been found to correlate with folding rates [210, 211, 212, 213, 214, 215]. In particular, two measures have emerged as being consistently well correlated: the ACO and the Long Range Order (LRO). LRO was found not only to correlate well overall [210] but also to correlate well across different structural classes of proteins [205]. Recently, the correlation of folding rates with ACO was explicitly derived from theory [216].

In contrast, relatively little work on the relationship between unfolding rates and native structure has been reported, with the existing studies suggesting that while native structure should correlate with unfolding [217], the measures of structural complexity that work well for predicting folding do not perform well for unfolding [218, 219, 220]. In particular, work by Jung *et al.* concluded that while structural complexity does correlate with both unfolding and folding rates, the predictive parameters are different [218, 219]. Harihar and Selvaraj compared LRO with unfolding rates, finding a moderate correlation overall, but greatly differing correlations for the alpha, beta, and mixed structural classes [220]. In these studies, relatively small datasets of ∼25 two-state folding proteins were used. In particular, the approach in two later studies was to use the small consensus dataset compiled by Maxwell *et al.* [221] in order to avoid noise in the data resulting from experimental differences [218, 220]. However, noise resulting from sequence specific effects can also be considerable, as noted in studies of homologous proteins [222, 223, 224]. Thus, the use of a small dataset for examining a relationship with such considerable noise may be a fraught endeavour. For instance, the original very strong correlation of folding rate with RCO [197] was later found to be considerably weaker when larger datasets including multi-state folders were used [198, 205].

Here, in order to clearly identify general relationships between folding/unfolding rates and structural complexity we have used a relatively large set of kinetic data for monomeric proteins obtained using similar experimental approaches. Proteins with disulfide bonds or large prosthetic groups are excluded because they are known to cause anomalous kinetics. The dataset is largely similar to that of Garbuzynskiy *et al.*, who recently elucidated relationships between protein length, stability and folding rates [225]. Using our dataset of 108 two- and multi-state folders (see Methods, Supplemental Table 4.4), we tested various measures of structural complexity, discovering that two commonly used parameters, ACO and LRO, not only correlate strongly with folding rates but also correlate strongly, and equally well, with unfolding rates. Furthermore, the results are very similar for different structural classes of proteins. Importantly, these results address the previously reported apparent differences between the structural determinants of unfolding and folding rates which may have been a consequence of the comparatively smaller datasets that obscured the true relationships. That the same measures of structural complexity are equally predictive of unfolding and folding rates has important implications for fundamental understanding of the process of protein unfolding. It also suggests that for the protein engineer, a key choice needs to be made when selecting a scaffold for design in order to achieve the desired balance between the typically desirable properties of fast folding and slow unfolding.

4.4 Results and Discussion

In order to identify general trends with increased confidence, we used a previously established large dataset [221, 225] augmented with additional proteins (see Methods). We analyzed this dataset using a range of measures of structural complexity found previously to be correlated with folding rates (Supplemental Table 4.2). Two well established parameters, ACO and LRO (see Methods), exhibited superior correlations, which are described in detail below. We note that the trends for unfolding rates for the full dataset shown in Figure 4.1 also hold in general (with some variations in statistical significance) for various subsets of the data (Table 4.1).

4.4.1 Unfolding rates correlate with ACO, LRO, folding rates, and stability

Strikingly, the logarithm of both k_u and k_f have equally negative correlation coefficients with ACO and LRO (Figure 4.1A-D, Table 4.1), which suggests these measures of structural complexity are similarly predictive of both rates (small variations in correlation are expected owing to differing experimental conditions and the necessary extrapolation of the rates from kinetic data). Thus, both unfolding and folding rates decrease with increasing structural complexity. These results contrast with those of previous studies using smaller datasets of 22 and 25 proteins which concluded that one parameter was not equally well suited for predicting both unfolding and folding [218, 219]. Further, we find that the correlation for unfolding holds well across different structural classes (Table 4.1), whereas another analysis [220] suggested all-beta proteins have a much weaker correlation with

LRO that is opposite in sign to that for all-alpha and mixed structural classes. These apparent discrepancies are likely caused by the small dataset sizes used in the earlier study where the alpha and beta classes had 5 and 7, compared here with 33 and 34 proteins, respectively.

We also note that unfolding rates are strongly correlated with folding rates (Figure 4.1E and Table 4.1), and that ACO and LRO are strongly correlated with rates at the transition midpoints (i.e. under conditions of equal thermostability) (Supplemental Figure 4.2, Supplemental Table 4.2). Thus, ACO and LRO may report on the structural complexity and relative energy of the transition state [226]. There is also a weaker correlation between protein length and the unfolding and folding rates (Supplemental Table 4.2). Together these trends indicate that protein topological complexity and size affect both folding and unfolding rates.

Lastly, there is a strong correlation between unfolding rate and thermodynamic stability (Figure 4.1F, Table 4.1). In contrast, the correlation of folding rate with thermodynamic stability is weak (Figure 4.1G, Table 4.1) as has been found previously [197, 227]. The larger contribution of unfolding rate to thermodynamic stability has been noted before [218, 219, 223, 224, 228], and is also apparent in the differences between the upper and lower limits of the dataset, where the upper limit on fast folding and unfolding is similar, while the lower bound for unfolding is substantially slower than that for folding (Supplemental Table 4.3). These results suggest that variations in thermodynamic stability, which are determined by the ratios of folding to unfolding rates, are dominated by unfolding rather than folding rates (Table 4.1). Why is this so? Folding may have a biologically imposed lower limit *in vivo*, such that it is sufficiently fast to avoid degradation or aggregation [229], and an upper limit imposed by physical constraints even for the most topologically simple folds [225]. Conversely, while there may be a similar physical limit for fast unfolding, the biological limit for slow unfolding, which may be related to the need for eventual protein turnover [230], may be more malleable due to the greatly differing roles and lifetimes of natural proteins.

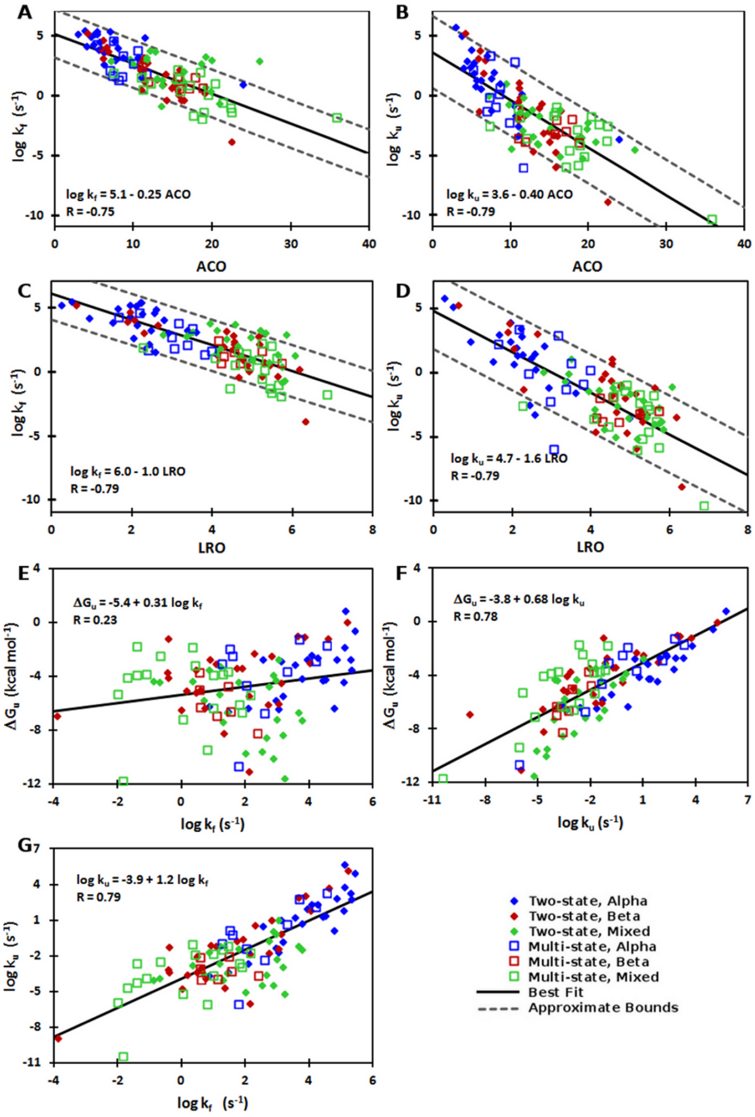


Figure 4.1: Correlations between structural complexity, folding and unfolding rates, and thermodynamic stability. Correlations are shown between (A) folding rates and ACO, (B) unfolding rates and ACO, (C) folding rates and LRO, (D) unfolding rates and LRO, (E) folding rates and thermodynamic stability, (F) unfolding rates and thermodynamic stability, and (G) unfolding and folding rates. The lines of best fit (solid black) and corresponding equations and correlation values are given for the whole dataset, values for subsets of data are given in Table 4.1 for two-state (filled diamonds), multistate (open squares), alpha (blue), beta (red), and mixed (green) proteins. Dotted lines for panels A–D denote ±10-fold and ±100-fold variation in k_f and k_u, respectively.

Table 4.1: Correlations and linear fits of unfolding and folding rate constants, measures of native structure, and thermodynamic stability

			Linear fit: $y = b + m * x$		Pearson correlation	
Parameter x	Parameter y	Dataset (size)	m	b	R	P^a
ACO	log k_f	Full (108)	-0.25	5.1	-0.75	$1.4 \text{x} 10^{-20}$
ACO	log k_f	Two-State (73)	-0.25	5.4	-0.73	$3.6 \text{x} 10^{-13}$
ACO	log k_f	Multi-State (35)	-0.20	4.0	-0.75	$2.7 \text{x} 10^{-7}$
ACO	log k_f	Alpha (33)	-0.22	5.3	-0.62	$1.1 \text{x} 10^{-4}$
ACO	log k_f	Beta (34)	-0.35	6.1	-0.86	$4.9 \text{x} 10^{-11}$
ACO	log k_f	Mixed (41)	-0.16	3.7	-0.52	$5.1 \text{x} 10^{-4}$
ACO	log k_f	Maxwell (28)	-0.14	4.1	-0.48	$9.5 \text{x} 10^{-3}$
ACO	log k_u	Full (108)	-0.40	3.6	-0.79	$3.0 \text{x} 10^{-24}$
ACO	log k_u	Two-State (73)	-0.46	4.4	-0.82	$3.9 \text{x} 10^{-19}$
ACO	log k_u	Multi-State (35)	-0.30	1.9	-0.70	$2.8 \text{x} 10^{-6}$
ACO	log k_u	Alpha (33)	-0.48	4.5	-0.70	$6.4 \text{x} 10^{-6}$
ACO	log k_u	Beta (34)	-0.55	5.4	-0.82	$3.5 \text{x} 10^{-9}$
ACO	log k_u	Mixed (41)	-0.25	1.0	-0.67	$2.0 \text{x} 10^{-6}$
ACO	log k_u	Maxwell (28)	-0.30	2.1	-0.71	$2.7 \text{x} 10^{-5}$
LRO	log k_f	Full (108)	-1.0	6.0	-0.79	$2.9 \text{x} 10^{-24}$
LRO	log k_f	Two-State (73)	-0.94	6.0	-0.80	$2.0 \text{x} 10^{-17}$
LRO	log k_f	Multi-State (35)	-1.0	5.6	-0.75	$2.7 \text{x} 10^{-7}$
LRO	log k_f	Alpha (33)	-0.96	5.9	-0.69	$1.0 \text{x} 10^{-5}$
LRO	log k_f	Beta (34)	-1.1	6.3	-0.82	$4.3 \text{x} 10^{-9}$
LRO	log k_f	Mixed (41)	-0.95	5.8	-0.54	$2.6 \text{x} 10^{-4}$
LRO	log k_f	Maxwell (28)	-0.93	6.2	-0.68	$7.6 \text{x} 10^{-5}$
LRO	log k_u	Full (108)	-1.6	4.7	-0.79	$6.9 \text{x} 10^{-24}$
LRO	log k_u	Two-State (73)	-1.6	4.9	-0.82	$3.4 \text{x} 10^{-19}$
LRO	log k_u	Multi-State (35)	-1.4	3.7	-0.64	$3.5 \text{x} 10^{-5}$

LRO	log k_u	Alpha (33)	-1.8	5.2	-0.68	$1.3x10^{-5}$
LRO	log k_u	Beta (34)	-1.7	5.9	-0.79	$3.6x10^{-8}$
LRO	log k_u	Mixed (41)	-1.2	3.0	-0.56	$1.4x10^{-4}$
LRO	log k_u	Maxwell (28)	-1.5	4.7	-0.74	$5.6x10^{-6}$
log k_f	ΔG_{F-U}	Full (108)	0.31	-5.4	0.23	$1.8x10^{-2}$
log k_f	ΔG_{F-U}	Two-State (73)	0.37	-5.6	0.27	$2.3x10^{-2}$
log k_f	ΔG_{F-U}	Multi-State (35)	0.23	-5.1	0.15	$4.0x10^{-1}$[NS]
log k_f	ΔG_{F-U}	Alpha (33)	0.78	-6.6	0.47	$5.2x10^{-3}$
log k_f	ΔG_{F-U}	Beta (34)	0.49	-5.3	0.36	$3.8x10^{-2}$
log k_f	ΔG_{F-U}	Mixed (41)	-0.52	-5.1	-0.33	$3.6x10^{-2}$
log k_f	ΔG_{F-U}	Maxwell (28)	-0.49	-4.4	-0.26	$1.9x10^{-1}$[NS]
log k_u	ΔG_{F-U}	Full (108)	0.68	-3.8	0.78	$2.4x10^{-23}$
log k_u	ΔG_{F-U}	Two-State (73)	0.69	-4.0	0.80	$2.2x10^{-17}$
log k_u	ΔG_{F-U}	Multi-State (35)	0.76	-3.0	0.79	$1.4x10^{-8}$
log k_u	ΔG_{F-U}	Alpha (33)	0.77	-4.4	0.89	$7.0x10^{-12}$
log k_u	ΔG_{F-U}	Beta (34)	0.69	-3.3	0.83	$1.7x10^{-9}$
log k_u	ΔG_{F-U}	Mixed (41)	0.82	-3.2	0.66	$3.0x10^{-6}$
log k_u	ΔG_{F-U}	Maxwell (28)	0.98	-3.8	0.77	$1.9x10^{-6}$
log k_f	log k_u	Full (108)	1.2	-3.9	0.79	$6.4x10^{-24}$
log k_f	log k_u	Two-State (73)	1.3	-4.1	0.79	$7.4x10^{-17}$
log k_f	log k_u	Multi-State (35)	1.2	-3.8	0.72	$1.1x10^{-6}$
log k_f	log k_u	Alpha (33)	1.6	-4.8	0.83	$2.6x10^{-9}$
log k_f	log k_u	Beta (34)	1.4	-3.9	0.82	$2.6x10^{-9}$
log k_f	log k_u	Mixed (41)	0.62	-3.7	0.50	$9.8x10^{-4}$
log k_f	log k_u	Maxwell (28)	0.64	-3.3	0.42	$2.5x10^{-2}$

Individual correlations and linear fits are shown for subsets of the data as in Figure 4.1. Additionally, values for the commonly used dataset of Maxwell *et al.* [221] are shown for comparison

[a] Two-tailed probability value

[NS] Correlation is not significant at the 0.05 level ($5.0 \text{ x } 10^{-2}$)

$\Delta G_{F-U} = G_F - G_U = -RT \ln(k_f/k_u)$, where R is the gas constant and T is the absolute temperature in Kelvin, gives the Gibbs free energy of the folded state relative to the unfolded state

4.4.2 Experimental and theoretical support for correlation between unfolding rates and native structure

Multiple lines of evidence suggest that unfolding rates should correlate with native structural complexity. First, while the relationships between unfolding rates and structure observed here may appear to be at odds with prior studies [218, 219, 220], this is likely spurious due to trends being obscured previously when analyzing smaller datasets with substantial noise. Specifically, for a given value of a structural parameter (ACO or LRO), the variation in the observed unfolding rates is ± ~10-fold larger than that for the folding rates (Figure 4.1A-D). Thus, compared to folding rates, to detect significant correlations between unfolding rates and structural parameters, the absolute range of unfolding rates needs to be larger. The smaller datasets used in previous unfolding analyses [218, 219, 220], which were based on the more curated set of Maxwell *et al.* [221], had a range of ~8 orders of magnitude for the unfolding rates ($6\text{x}10^{-6}$ to $1\text{x}10^{2}$ s^{-1}). The larger dataset used here spans ~16 orders of magnitude ($4\text{x}10^{-11}$ to $5\text{x}10^{5}$ s^{-1}), and as such the correlation between unfolding rate and structural complexity can be observed more clearly. Second, as the rates at the transition midpoint (where k_u is equal to k_f, and ΔG is 0) report on the transition state energy, the correlation of these rates with measures of structural complexity suggests that both the folding and unfolding rates (under conditions of different ΔG) should also be correlated with those same measures of structural complexity. Third, a recently developed method based on physical principles and protein structural class and size was able to predict both unfolding and folding rates for a set of 52 two-state folding proteins [231]. Finally, an analysis of 53 two- and 19 multi-state folders using a complex fractal parameter found comparable correlations with unfolding and folding rates, although the strength of the correlation was weaker than reported here [215]. The above considerations provide support for our observation of the significant correlations of structural complexity with both folding and unfolding rates.

4.4.3 Implications for design

The correlations reported herein indicate that the same measures of structural complexity predict both folding and unfolding rates equally well and, consequently, it may be difficult to modulate one aspect of the structure to alter (e.g. gain) folding speed, while leaving unfolding speed unaffected. Thus, it may seem a daunting task to achieve the desirable outcome of both fast folding and slow unfolding simultaneously. However, while the correlations of structural complexity and folding/unfolding rates have high statistical significance (Table 4.1), there is nevertheless considerable variation around the lines of best fit, which we roughly estimate to be in the range of ± two orders of magnitude for folding, and ± three orders of magnitude for unfolding rates (Figure 4.1A-D). Although some of this variation may be caused by more complex topological features such as nested structures [216, 226, 232], it has been noted previously [205, 233] and well documented experimentally, for example by comparison of homologous proteins [222, 223, 224, 227, 233], that while the native structure may place upper and lower boundaries on folding and unfolding rates, sequence-specific effects can be substantial. This is also illustrated by the effects, sometimes quite large, of point mutations on kinetics [234]. In addition, single mutations tend to have a larger absolute effect on unfolding rather than folding rates based on our analysis of a dataset collected by Naganathan and Muñoz [235] where the change in unfolding rate is ∼15-fold greater on average than for folding rates (Supplemental Table 4.4). Together the above points suggest that while the scaffold may define broad ranges for folding/unfolding rates, sequence specific engineering can provide substantial scope to modulate these rates in order to achieve to some extent, fast folding and slow unfolding. Fortunately, much work has been done on the sequence-specific determinants of folding and unfolding rates, and some lessons may be learned from this. First, the nature of functional sites in proteins may modulate topological complexity and alter kinetics. This was studied for two beta-trefoil proteins: the functional myristoyl binding site of Hisactophilin is a cavity within the protein core which reduces structural complexity and so may speed folding and unfolding, whereas the binding site of Interleukin-1 beta is formed by two long loops which increase structural complexity and may slow the kinetics [236]. Second, both residual structure in the denatured state [237, 238], and non-native interactions in the transition state [239, 240, 241] can increase the folding rate independent of the overall topology. Third, unfolding rates can be slowed by introducing hydrophobic residues on the surface of the native structure, which may increase local rigidity and the barrier to local hydration [242], or by surface electrostatic interactions, which may act as staples [243]. Lastly, it may be possible through computational simulation to identify particular weak points in the structure which, if strengthened, could increase unfolding cooperativity and therefore increase the height of the unfolding barrier [244]. Thus, multiple approaches may be used to modulate sequence-specific interactions in order to alter folding/unfolding rates.

4.5 Conclusions

We have shown, using a large dataset with highly significant correlations, that the measures of structural complexity that have emerged as strong predictors of folding rates, ACO and LRO, are equally predictive of unfolding rates, contrary to what has been reported previously when using smaller datasets [218, 219, 220]. In addition, the correlations are fairly robust to kinetic mechanism, whether two- or multi-state, and structural class, whether alpha, beta, or mixed. From a fundamental protein folding point of view, this suggests that the structural complexity reported on by ACO and LRO is a key determinant of both folding and unfolding processes.

These results have important implications for protein engineering and design. Specifically, a topologically simple scaffold may fold quickly, and so attain *in vivo* activity [225, 229]; but, it will also unfold quickly, reducing kinetic stability and resistance to thermal or chemical denaturation and degradation by proteases [245, 246]. On the other hand, a topologically complex scaffold may possess the high kinetic stability that would be ideal for harsh industrial conditions or crowded and proteolytic biological environments [246], but it may have difficulty folding fast enough to be biologically viable [199, 225]. These conflicting kinetic constraints may limit the prospects when both fast folding and kinetic stability are required, or when improving a particular scaffold that is desirable for other reasons (such as function). However, these difficulties can be overcome as in existing proteins where the large variation in folding/unfolding rates and significant effects of point mutations [247] Supplemental Table 4.4) demonstrate that appropriately designed sequences can ease constraints placed on the folding and unfolding rates by the structural complexity of the protein scaffold.

4.6 Methods

Our dataset is largely that of Garbuzynskiy *et al.* [225] (which includes the smaller dataset of Maxwell *et al.* [221]). We added data from the Kinetic DataBase [248], and other published kinetic data including our own [223, 249, 79, 250] as well as data on our engineered ThreeFoil protein [44] (for which the kinetic experiments will be published separately). In adding data we followed the general criterion used by Garbuzynskiy *et al.* [225], using only monomeric, single domain proteins, which lack disulfide bonds and prosthetic groups. In addition, the experimental temperature was in the range, or could be reliably extrapolated, to ∼25 °C, and the folding and unfolding rates were measured at, or could be extrapolated to, 0 M denaturant. The specific details of the sources for each member of the 108 protein dataset are given in Supplemental Table 4.3.

Absolute Contact Order (ACO), the average sequence separation between contacting heavy atoms, was calculated as described by Plaxco *et al.* and later by Ivankov *et al.* [197, 198]

$$ACO = \frac{1}{N_c} \sum_{i,j}^{N_c} |i - j| \tag{4.1}$$

where N_c is the total number of contacts between heavy atoms, and $|i-j|$ is the sequence separation in residues for a given contacting pair of atoms. Contacts are considered between heavy atoms less than 6 Å apart. Here we modify the original definition to only count contacts where $|i - j| > 1$, since all residues would otherwise be in contact with their covalently attached neighbours.

The formula for calculating Long Range Order (LRO) is that of Gromiha and Selvaraj [210]

$$LRO = \frac{1}{L} \sum_{i,j}^{R_c} \mathbf{n}_{i,j} \tag{4.2}$$

where L is the protein length, R_c is the total number of contacting residues and $\mathbf{n}_{i,j}$ is 1 when $|i-j| \geq 12$ and 0 otherwise. We have modified the definition of a contact between residues to be the same as in ACO rather than the original criterion of a Cα separation less than 8 Å. This modified form yields slightly improved correlations for both folding and unfolding; using 6 Å may correct for underestimation of long range contacts between large residues in cores.

4.7 Supplemental Information

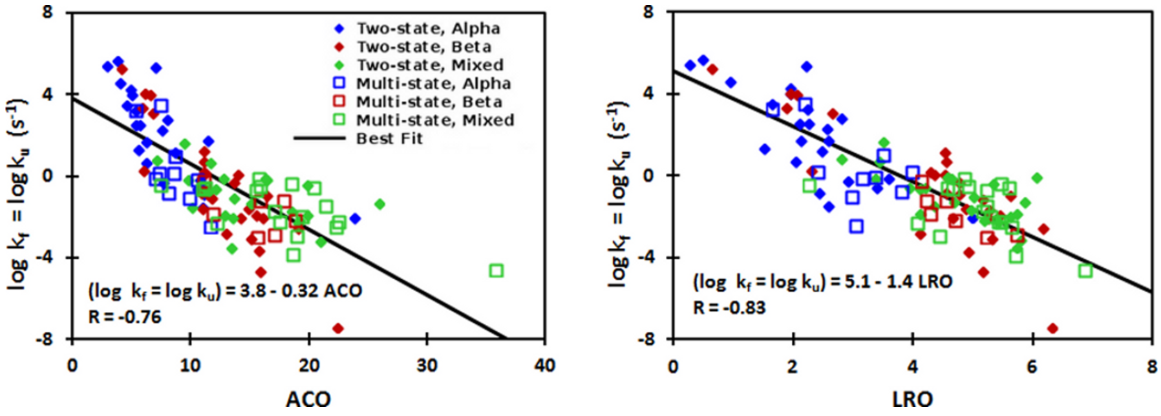


Figure 4.2: Correlations between structural complexity and the rates of folding and unfolding at the transition midpoint. Correlations are shown between for the folding/unfolding rate at the transition midpoint (log k$_f$ = log k$_u$) for A) ACO and B) LRO. Rates at the transition midpoint are used because the thermodynamic stability is zero allowing for the correction of effects owing to differing stabilities.

Table 4.2: Correlations and linear fits of unfolding and folding rate constants, measures of native structure, and thermodynamic stability, extended

			Linear fit: $y = b + m * x$		Pearson correlation	
Parameter x	Parameter y	Dataset (size)	m	b	R	P^a
ACO	log k$_f$ = log k$_u$	Full (108)	-0.32	3.8	-0.76	8.2x10^{-22}
ACO	log k$_f$ = log k$_u$	Two-State (73)	-0.40	4.7	-0.80	2.0x10^{-17}
ACO	log k$_f$ = log k$_u$	Multi-State (35)	-0.19	1.7	-0.70	2.3x10^{-6}
ACO	log k$_f$ = log k$_u$	Alpha (33)	-0.43	4.9	-0.71	3.4x10^{-6}
ACO	log k$_f$ = log k$_u$	Beta (34)	-0.52	5.9	-0.86	4.7x10^{-11}
ACO	log k$_f$ = log k$_u$	Mixed (41)	-0.14	1.0	-0.61	2.2x10^{-5}
ACO	log k$_f$ = log k$_u$	Maxwell (28)	-0.21	2.2	-0.69	4.6x10^{-5}
LRO	log k$_f$ = log k$_u$	Full (108)	-1.4	5.1	-0.83	1.9x10^{-28}
LRO	log k$_f$ = log k$_u$	Two-State (73)	-1.4	5.5	-0.85	1.8x10^{-21}
LRO	log k$_f$ = log k$_u$	Multi-State (35)	-0.98	3.2	-0.71	1.9x10^{-6}
LRO	log k$_f$ = log k$_u$	Alpha (33)	-1.8	6.0	-0.77	1.7x10^{-7}
LRO	log k$_f$ = log k$_u$	Beta (34)	-1.7	6.6	-0.85	1.3x10^{-10}
LRO	log k$_f$ = log k$_u$	Mixed (41)	-0.85	2.8	-0.61	2.3x10^{-5}
LRO	log k$_f$ = log k$_u$	Maxwell (28)	-1.2	4.6	-0.81	1.4x10^{-7}
RCO	log k$_f$	Full (108)	-3.9	2.5	-0.08	3.9x10^{-1NS}
RCO	log k$_f$	Two-State (73)	-16	4.9	-0.34	2.8x10^{-3}
RCO	log k$_f$	Multi-State (35)	-0.036	1.1	0.00	1.0x10^{0NS}
RCO	log k$_f$	Alpha (33)	27	0.7	0.51	2.4x10^{-3}
RCO	log k$_f$	Beta (34)	7.7	0.28	0.15	4.0x10^{-1NS}
RCO	log k$_f$	Mixed (41)	12	-0.55	0.27	8.8x10^{-2NS}
RCO	log k$_f$	Maxwell (28)	-19	5.3	-0.56	1.7x10^{-3}
RCO	log k$_u$	Full (108)	0.65	-1.6	0.01	9.3x10^{-1NS}
RCO	log k$_u$	Two-State (73)	-9.2	0.36	-0.12	3.1x10^{-1NS}
RCO	log k$_u$	Multi-State (35)	-1.9	-2.3	-0.02	9.0x10^{-1NS}

RCO	log k$_u$	Alpha (33)	51	-4.7	0.52	2.0x10^{-3}
RCO	log k$_u$	Beta (34)	39	-8.1	0.46	6.4x10^{-3}
RCO	log k$_u$	Mixed (41)	7.9	-4.1	0.14	3.8x10^{-1NS}
RCO	log k$_u$	Maxwell (28)	-14	0.3	-0.26	1.7x10^{-1NS}
RCO	log k$_f$ = log k$_u$	Full (108)	-2.2	0.024	-0.04	7.0x10^{-1NS}
RCO	log k$_f$ = log k$_u$	Two-State (73)	-12	1.8	-0.17	1.4x10^{-1NS}
RCO	log k$_f$ = log k$_u$	Multi-State (35)	-2.1	-0.87	-0.04	8.1x10^{-1NS}
RCO	log k$_f$ = log k$_u$	Alpha (33)	42	-2.8	0.48	4.6x10^{-3}
RCO	log k$_f$ = log k$_u$	Beta (34)	29	-5.5	0.38	2.7x10^{-2}
RCO	log k$_f$ = log k$_u$	Mixed (41)	4.8	-2.0	0.14	4.0x10^{-1NS}
RCO	log k$_f$ = log k$_u$	Maxwell (28)	-14	1.7	-0.39	3.9x10^{-2}
TCD	log k$_f$	Full (108)	-2.8	4.8	-0.37	7.5x10^{-5}
TCD	log k$_f$	Two-State (73)	-5.0	7.6	-0.65	5.6x10^{-10}
TCD	log k$_f$	Multi-State (35)	-1.4	2.4	-0.21	2.3x10^{-1NS}
TCD	log k$_f$	Alpha (33)	2.0	2.0	0.19	2.9x10^{-1NS}
TCD	log k$_f$	Beta (34)	-3.9	5.8	-0.42	1.2x10^{-2}
TCD	log k$_f$	Mixed (41)	0.99	0.085	0.14	3.9x10^{-1NS}
TCD	log k$_f$	Maxwell (28)	-3.6	6.4	-0.60	6.5x10^{-4}
TCD	log k$_u$	Full (108)	-4.1	2.6	-0.35	2.4x10^{-4}
TCD	log k$_u$	Two-State (73)	-6.7	6.0	-0.55	6.1x10^{-7}
TCD	log k$_u$	Multi-State (35)	-2.3	-0.47	-0.20	2.4x10^{-1NS}
TCD	log k$_u$	Alpha (33)	1.4	-0.39	0.07	6.9x10^{-1NS}
TCD	log k$_u$	Beta (34)	-2.3	0.82	-0.15	3.8x10^{-1NS}
TCD	log k$_u$	Mixed (41)	-0.23	-2.8	-0.03	8.7x10^{-1NS}
TCD	log k$_u$	Maxwell (28)	-3.8	2.5	-0.42	2.5x10^{-2}
TCD	log k$_f$ = log k$_u$	Full (108)	-3.9	3.6	-0.40	1.6x10^{-5}
TCD	log k$_f$ = log k$_u$	Two-State (73)	-6.4	6.8	-0.59	4.2x10^{-8}
TCD	log k$_f$ = log k$_u$	Multi-State (35)	-1.7	0.43	-0.24	1.6x10^{-1NS}
TCD	log k$_f$ = log k$_u$	Alpha (33)	0.037	1.5	0.00	1.0x10^{0NS}
TCD	log k$_f$ = log k$_u$	Beta (34)	-3.8	3.4	-0.28	1.0x10^{-1NS}

TCD	log k$_f$ = log k$_u$	Mixed (41)	-0.17	-1.1	-0.03	8.5x10^{-1NS}
TCD	log k$_f$ = log k$_u$	Maxwell (28)	-3.5	3.4	-0.55	2.7x10^{-3}
SRO	log k$_f$	Full (108)	0.55	-1.7	0.35	2.1x10^{-4}
SRO	log k$_f$	Two-State (73)	0.66	-1.9	0.44	8.3x10^{-5}
SRO	log k$_f$	Multi-State (35)	0.62	-3.2	0.40	1.7x10^{-2}
SRO	log k$_f$	Alpha (33)	-0.85	10.	-0.37	3.4x10^{-2}
SRO	log k$_f$	Beta (34)	-0.38	3.6	-0.08	6.6x10^{-1NS}
SRO	log k$_f$	Mixed (41)	-0.33	3.3	-0.11	5.0x10^{-1NS}
SRO	log k$_f$	Maxwell (28)	0.59	-1.6	0.50	7.4x10^{-3}
SRO	log k$_u$	Full (108)	0.68	-6.0	0.27	4.0x10^{-3}
SRO	log k$_u$	Two-State (73)	0.68	-5.5	0.28	1.5x10^{-2}
SRO	log k$_u$	Multi-State (35)	1.1	-10.	0.43	9.1x10^{-3}
SRO	log k$_u$	Alpha (33)	-2.5	21	-0.57	5.7x10^{-3}
SRO	log k$_u$	Beta (34)	-2.1	9.8	-0.26	1.4x10^{-1NS}
SRO	log k$_u$	Mixed (41)	0.83	-8.4	0.22	1.7x10^{-1NS}
SRO	log k$_u$	Maxwell (28)	0.71	-6.5	0.40	3.6x10^{-2}
SRO	log k$_f$ = log k$_u$	Full (108)	0.63	-4.5	0.31	1.0x10^{-3}
SRO	log k$_f$ = log k$_u$	Two-State (73)	0.73	-4.7	0.35	2.6x10^{-3}
SRO	log k$_f$ = log k$_u$	Multi-State (35)	0.67	-5.8	0.43	1.0x10^{-2}
SRO	log k$_f$ = log k$_u$	Alpha (33)	-2.0	18	-0.53	1.5x10^{-3}
SRO	log k$_f$ = log k$_u$	Beta (34)	-1.9	9.5	-0.26	1.4x10^{-1NS}
SRO	log k$_f$ = log k$_u$	Mixed (41)	0.56	-4.9	0.23	1.4x10^{-1NS}
SRO	log k$_f$ = log k$_u$	Maxwell (28)	0.63	-4.7	0.49	8.0x10^{-3}
NonLocalCO	log k$_f$	Full (108)	-0.17	5.2	-0.59	1.1x10^{-11}
NonLocalCO	log k$_f$	Two-State (73)	-0.13	4.8	-0.46	4.3x10^{-5}
NonLocalCO	log k$_f$	Multi-State (35)	-0.19	5.3	-0.74	3.6x10^{-7}
NonLocalCO	log k$_f$	Alpha (33)	-0.11	5.4	-0.58	4.5x10^{-4}
NonLocalCO	log k$_f$	Beta (34)	-0.27	6.2	-0.77	7.7x10^{-8}
NonLocalCO	log k$_f$	Mixed (41)	-0.14	4.4	-0.53	3.3x10^{-4}
NonLocalCO	log k$_f$	Maxwell (28)	-0.03	2.9	-0.14	4.7x10^{-1NS}

NonLocalCO	log k$_u$	Full (108)	-0.31	4.5	-0.71	7.1x10^{-18}
NonLocalCO	log k$_u$	Two-State (73)	-0.32	4.7	-0.69	2.0x10^{-11}
NonLocalCO	log k$_u$	Multi-State (35)	-0.28	3.9	-0.70	3.5x10^{-6}
NonLocalCO	log k$_u$	Alpha (33)	-0.27	5.3	-0.74	6.6x10^{-7}
NonLocalCO	log k$_u$	Beta (34)	-0.45	6.0	-0.79	3.6x10^{-8}
NonLocalCO	log k$_u$	Mixed (41)	-0.22	2.0	-0.65	3.4x10^{-6}
NonLocalCO	log k$_u$	Maxwell (28)	-0.14	0.98	-0.45	1.6x10^{-2}
NonLocalCO	log k$_f$ = log k$_u$	Full (108)	-0.23	4.3	-0.65	1.5x10^{-14}
NonLocalCO	log k$_f$ = log k$_u$	Two-State (73)	-0.25	4.7	-0.62	4.5x10^{-9}
NonLocalCO	log k$_f$ = log k$_u$	Multi-State (35)	-0.18	3.0	-0.70	2.3x10^{-6}
NonLocalCO	log k$_f$ = log k$_u$	Alpha (33)	-0.22	5.4	-0.70	5.0x10^{-6}
NonLocalCO	log k$_f$ = log k$_u$	Beta (34)	-0.42	6.4	-0.81	5.3x10^{-9}
NonLocalCO	log k$_f$ = log k$_u$	Mixed (41)	-0.12	1.5	-0.59	5.1x10^{-5}
NonLocalCO	log k$_f$ = log k$_u$	Maxwell (28)	-0.085	1.2	-0.38	4.6x10^{-2}
LocalCO	log k$_f$	Full (108)	0.28	0.96	0.56	4.5x10^{-10}
LocalCO	log k$_f$	Two-State (73)	0.24	1.5	0.57	1.7x10^{-7}
LocalCO	log k$_f$	Multi-State (35)	0.58	-0.70	0.60	1.4x10^{-4}
LocalCO	log k$_f$	Alpha (33)	0.12	2.6	0.44	1.1x10^{-2}
LocalCO	log k$_f$	Beta (34)	0.74	-0.077	0.66	2.5x10^{-5}
LocalCO	log k$_f$	Mixed (41)	0.58	-0.12	0.28	8.1x10^{-2NS}
LocalCO	log k$_f$	Maxwell (28)	0.50	0.91	0.62	4.5x10^{-4}
LocalCO	log k$_u$	Full (108)	0.47	-3.3	0.60	4.9x10^{-12}
LocalCO	log k$_u$	Two-State (73)	0.43	-2.7	0.64	1.6x10^{-9}
LocalCO	log k$_u$	Multi-State (35)	0.84	-5.1	0.54	8.8x10^{-4}
LocalCO	log k$_u$	Alpha (33)	0.27	-1.3	0.54	1.3x10^{-3}
LocalCO	log k$_u$	Beta (34)	1.2	-4.5	0.66	2.1x10^{-5}
LocalCO	log k$_u$	Mixed (41)	0.99	-5.2	0.38	1.5x10^{-2}
LocalCO	log k$_u$	Maxwell (28)	0.78	-3.9	0.63	2.9x10^{-3}
LocalCO	log k$_f$ = log k$_u$	Full (108)	0.42	-1.8	0.65	4.3x10^{-14}
LocalCO	log k$_f$ = log k$_u$	Two-State (73)	0.39	-1.5	0.65	6.0x10^{-10}

LocalCO	log k$_f$ = log k$_u$	Multi-State (35)	0.63	-3.0	0.64	3.9x10^{-5}
LocalCO	log k$_f$ = log k$_u$	Alpha (33)	0.26	-0.32	0.57	4.7x10^{-4}
LocalCO	log k$_f$ = log k$_u$	Beta (34)	1.1	-3.3	0.68	8.3x10^{-6}
LocalCO	log k$_f$ = log k$_u$	Mixed (41)	0.72	-2.9	0.43	4.5x10^{-3}
LocalCO	log k$_f$ = log k$_u$	Maxwell (28)	0.62	-2.2	0.71	2.4x10^{-5}
Nα	log k$_f$	Full (108)	-0.048	4.3	-0.64	7.3x10^{-14}
Nα	log k$_f$	Two-State (73)	-0.042	4.1	-0.47	2.4x10^{-5}
Nα	log k$_f$	Multi-State (35)	-0.050	4.3	-0.80	9.0x10^{-9}
Nα	log k$_f$	Alpha (33)	-0.037	5.0	-0.72	2.2x10^{-6}
Nα	log k$_f$	Beta (34)	-0.076	4.6	-0.72	1.8x10^{-6}
Nα	log k$_f$	Mixed (41)	-0.041	3.6	-0.66	3.3x10^{-6}
Nα	log k$_f$	Maxwell (28)	-0.0089	2.7	-0.17	3.9x10^{-1NS}
Nα	log k$_u$	Full (108)	-0.078	2.2	-0.68	8.4x10^{-16}
Nα	log k$_u$	Two-State (73)	-0.094	2.7	-0.65	3.3x10^{-10}
Nα	log k$_u$	Multi-State (35)	-0.070	2.0	-0.69	5.6x10^{-6}
Nα	log k$_u$	Alpha (33)	-0.070	3.5	-0.72	1.8x10^{-6}
Nα	log k$_u$	Beta (34)	-0.15	4.1	-0.84	4.1x10^{-10}
Nα	log k$_u$	Mixed (41)	-0.047	-0.26	-0.60	3.5x10^{-5}
Nα	log k$_u$	Maxwell (28)	-0.031	-0.46	-0.39	4.0x10^{-2}
Nα	log k$_f$ = log k$_u$	Full (108)	-0.060	2.6	-0.63	1.8x10^{-13}
Nα	log k$_f$ = log k$_u$	Two-State (73)	-0.077	3.1	-0.61	1.1x10^{-8}
Nα	log k$_f$ = log k$_u$	Multi-State (35)	-0.046	1.8	-0.71	1.9x10^{-6}
Nα	log k$_f$ = log k$_u$	Alpha (33)	-0.060	4.0	-0.71	3.3x10^{-6}
Nα	log k$_f$ = log k$_u$	Beta (34)	-0.13	4.5	-0.84	4.2x10^{-10}
Nα	log k$_f$ = log k$_u$	Mixed (41)	-0.029	0.39	-0.58	6.1x10^{-5}
Nα	log k$_f$ = log k$_u$	Maxwell (28)	-0.020	0.33	-0.35	6.7x10^{-2NS}
CC	log k$_f$	Full (108)	32	-16	0.69	1.7x10^{-16}
CC	log k$_f$	Two-State (73)	27	-13	0.61	9.5x10^{-9}
CC	log k$_f$	Multi-State (35)	59	-32	0.83	9.9x10^{-10}
CC	log k$_f$	Alpha (33)	25	-11	0.73	1.2x10^{-6}

CC	log k_f	Beta (34)	28	-15	0.73	$1.2x10^{-6}$
CC	log k_f	Mixed (41)	31	-16	0.51	$5.9x10^{-4}$
CC	log k_f	Maxwell (28)	2.8	0.69	0.05	$7.9x10^{-1NS}$
CC	log k_u	Full (108)	57	-34	0.78	$3.3x10^{-23}$
CC	log k_u	Two-State (73)	56	-34	0.80	$2.4x10^{-17}$
CC	log k_u	Multi-State (35)	83	-48	0.71	$1.6x10^{-6}$
CC	log k_u	Alpha (33)	49	-28	0.75	$5.2x10^{-7}$
CC	log k_u	Beta (34)	53	-32	0.83	$1.1x10^{-9}$
CC	log k_u	Mixed (41)	44	-28	0.58	$6.9x10^{-5}$
CC	log k_u	Maxwell (28)	38	-23	0.48	$9.9x10^{-3}$
CC	log k_f = log k_u	Full (108)	47	-27	0.78	$2.4x10^{-23}$
CC	log k_f = log k_u	Two-State (73)	48	-28	0.77	$2.5x10^{-15}$
CC	log k_f = log k_u	Multi-State (35)	57	-33	0.79	$2.0x10^{-8}$
CC	log k_f = log k_u	Alpha (33)	45	-25	0.78	$7.1x10^{-8}$
CC	log k_f = log k_u	Beta (34)	47	-28	0.83	$1.0x10^{-9}$
CC	log k_f = log k_u	Mixed (41)	26	-16	0.54	$2.9x10^{-4}$
CC	log k_f = log k_u	Maxwell (28)	18	-11	0.32	$9.4x10^{-2NS}$
L	log k_f	Full (108)	-0.024	4.3	-0.64	$1.2x10^{-13}$
L	log k_f	Two-State (73)	-0.021	4.1	-0.46	$3.5x10^{-5}$
L	log k_f	Multi-State (35)	-0.025	4.4	-0.80	$7.6x10^{-9}$
L	log k_f	Alpha (33)	-0.018	5.0	-0.72	$2.8x10^{-6}$
L	log k_f	Beta (34)	-0.037	4.6	-0.70	$3.4x10^{-6}$
L	log k_f	Mixed (41)	-0.020	3.6	-0.64	$6.4x10^{-6}$
L	log k_f	Maxwell (28)	-0.0040	2.6	-0.16	$4.3x10^{-1NS}$
L	log k_u	Full (108)	-0.093	2.3	-0.68	$6.7x10^{-16}$
L	log k_u	Two-State (73)	-0.046	2.7	-0.65	$4.2x10^{-10}$
L	log k_u	Multi-State (35)	-0.036	2.2	-0.69	$3.6x10^{-6}$
L	log k_u	Alpha (33)	-0.035	3.6	-0.72	$2.4x10^{-6}$
L	log k_u	Beta (34)	-0.073	4.3	-0.84	$6.2x10^{-10}$
L	log k_u	Mixed (41)	-0.023	-0.21	-0.60	$3.5x10^{-5}$

L	log k_u	Maxwell (28)	-0.015	-0.44	-0.39	3.8x10^{-2}
L	log k_f = log k_u	Full (108)	-0.030	2.6	-0.64	1.5x10^{-13}
L	log k_f = log k_u	Two-State (73)	-0.038	3.2	-0.61	1.2x10^{-8}
L	log k_f = log k_u	Multi-State (35)	-0.023	1.9	-0.71	1.5x10^{-6}
L	log k_f = log k_u	Alpha (33)	-0.030	4.0	-0.71	3.8x10^{-6}
L	log k_f = log k_u	Beta (34)	-0.065	4.6	-0.83	9.7x10^{-10}
L	log k_f = log k_u	Mixed (41)	-0.014	0.43	-0.58	6.4x10^{-5}
L	log k_f = log k_u	Maxwell (28)	-0.0098	0.34	-0.35	6.5x10^{-2NS}
L	ΔG_{F-U}	Full (108)	-0.021	-2.7	-0.42	4.6x10^{-6}
L	ΔG_{F-U}	Two-State (73)	-0.035	-1.9	-0.57	1.2x10^{-7}
L	ΔG_{F-U}	Multi-State (35)	-0.014	-3.0	-0.29	9.5x10^{-2NS}
L	ΔG_{F-U}	Alpha (33)	-0.023	-2.0	-0.54	1.2x10^{-3}
L	ΔG_{F-U}	Beta (34)	-0.049	-0.42	-0.68	1.1x10^{-5}
L	ΔG_{F-U}	Mixed (41)	-0.0047	-5.1	-0.10	5.5x10^{-1NS}
L	ΔG_{F-U}	Maxwell (28)	-0.015	-4.2	-0.31	1.1x10^{-1NS}
$L^{2/3}$	log k_f	Full (108)	-0.18	5.7	-0.67	4.1x10^{-15}
$L^{2/3}$	log k_f	Two-State (73)	-0.16	5.3	-0.51	3.5x10^{-6}
$L^{2/3}$	log k_f	Multi-State (35)	-0.20	6.2	-0.81	3.0x10^{-9}
$L^{2/3}$	log k_f	Alpha (33)	-0.14	6.0	-0.75	6.4x10^{-7}
$L^{2/3}$	log k_f	Beta (34)	-0.24	6.1	-0.72	1.4x10^{-6}
$L^{2/3}$	log k_f	Mixed (41)	-0.15	4.8	-0.64	6.8x10^{-6}
$L^{2/3}$	log k_f	Maxwell (28)	-0.028	2.8	-0.14	4.9x10^{-1NS}
$L^{2/3}$	log k_u	Full (108)	-0.30	4.7	-0.72	1.3x10^{-18}
$L^{2/3}$	log k_u	Two-State (73)	-0.36	5.5	-0.72	8.3x10^{-13}
$L^{2/3}$	log k_u	Multi-State (35)	-0.28	4.7	-0.71	2.1x10^{-6}
$L^{2/3}$	log k_u	Alpha (33)	-0.27	5.7	-0.77	2.0x10^{-7}
$L^{2/3}$	log k_u	Beta (34)	-0.48	7.1	-0.86	1.0x10^{-10}
$L^{2/3}$	log k_u	Mixed (41)	-0.18	1.3	-0.60	3.4x10^{-5}
$L^{2/3}$	log k_u	Maxwell (28)	-0.13	0.74	-0.42	2.6x10^{-2}
$L^{2/3}$	log k_f = log k_u	Full (108)	-0.24	4.6	-0.69	2.6x10^{-16}

$L^{2/3}$	log k_f = log k_u	Two-State (73)	-0.30	5.5	-0.68	5.5x10^{-11}
$L^{2/3}$	log k_f = log k_u	Multi-State (35)	-0.19	3.6	-0.74	4.8x10^{-7}
$L^{2/3}$	log k_f = log k_u	Alpha (33)	-0.24	5.9	-0.76	2.3x10^{-7}
$L^{2/3}$	log k_f = log k_u	Beta (34)	-0.42	7.1	-0.85	1.7x10^{-10}
$L^{2/3}$	log k_f = log k_u	Mixed (41)	-0.11	1.4	-0.59	5.5x10^{-5}
$L^{2/3}$	log k_f = log k_u	Maxwell (28)	-0.081	1.0	-0.37	5.5x10^{-2NS}
$L^{2/3}$	ΔG_{F-U}	Full (108)	-0.17	-1.3	-0.46	4.2x10^{-7}
$L^{2/3}$	ΔG_{F-U}	Two-State (73)	-0.27	0.23	-0.63	2.4x10^{-9}
$L^{2/3}$	ΔG_{F-U}	Multi-State (35)	-0.11	-2.0	-0.29	8.8x10^{-2NS}
$L^{2/3}$	ΔG_{F-U}	Alpha (33)	-0.18	-0.53	-0.59	3.2x10^{-4}
$L^{2/3}$	ΔG_{F-U}	Beta (34)	-0.32	1.4	-0.69	6.2x10^{-6}
$L^{2/3}$	ΔG_{F-U}	Mixed (41)	-0.037	-4.8	-0.10	5.4x10^{-1NS}
$L^{2/3}$	ΔG_{F-U}	Maxwell (28)	-0.14	-2.8	-0.35	6.5x10^{-2NS}
$L^{1/2}$	log k_f	Full (108)	-0.52	7.0	-0.68	1.1x10^{-15}
$L^{1/2}$	log k_f	Two-State (73)	-0.46	6.5	-0.53	1.2x10^{-6}
$L^{1/2}$	log k_f	Multi-State (35)	-0.61	7.9	-0.82	2.1x10^{-9}
$L^{1/2}$	log k_f	Alpha (33)	-0.40	7.0	-0.75	3.9x10^{-7}
$L^{1/2}$	log k_f	Beta (34)	-0.66	7.5	-0.73	9.9x10^{-7}
$L^{1/2}$	log k_f	Mixed (41)	-0.46	6.1	-0.64	7.5x10^{-6}
$L^{1/2}$	log k_f	Maxwell (28)	-0.077	3.0	-0.12	5.3x10^{-1NS}
$L^{1/2}$	log k_u	Full (108)	-0.89	7.0	-0.74	6.0x10^{-20}
$L^{1/2}$	log k_u	Two-State (73)	-1.0	8.1	-0.75	4.1x10^{-14}
$L^{1/2}$	log k_u	Multi-State (35)	-0.86	7.2	-0.71	1.7x10^{-6}
$L^{1/2}$	log k_u	Alpha (33)	-0.80	7.7	-0.79	6.3x10^{-8}
$L^{1/2}$	log k_u	Beta (34)	-1.3	9.8	-0.86	4.7x10^{-11}
$L^{1/2}$	log k_u	Mixed (41)	-0.54	2.8	-0.60	3.5x10^{-5}
$L^{1/2}$	log k_u	Maxwell (28)	-0.40	1.9	-0.43	2.1x10^{-2}
$L^{1/2}$	log k_f = log k_u	Full (108)	-0.70	6.4	-0.71	9.2x10^{-18}
$L^{1/2}$	log k_f = log k_u	Two-State (73)	-0.87	7.7	-0.70	3.9x10^{-12}
$L^{1/2}$	log k_f = log k_u	Multi-State (35)	-0.57	5.3	-0.75	2.8x10^{-7}

L$^{1/2}$	log k$_f$ = log k$_u$	Alpha (33)	-0.70	7.7	-0.79	5.7x10^{-8}
L$^{1/2}$	log k$_f$ = log k$_u$	Beta (34)	-1.2	9.5	-0.86	7.8x10^{-11}
L$^{1/2}$	log k$_f$ = log k$_u$	Mixed (41)	-0.34	2.3	-0.59	5.3x10^{-5}
L$^{1/2}$	log k$_f$ = log k$_u$	Maxwell (28)	-0.25	1.8	-0.37	5.1x10^{-1NS}
L$^{1/2}$	ΔG$_{F-U}$	Full (108)	-0.50	0.054	-0.48	1.2x10^{-7}
L$^{1/2}$	ΔG$_{F-U}$	Two-State (73)	-0.79	2.2	-0.65	4.1x10^{-10}
L$^{1/2}$	ΔG$_{F-U}$	Multi-State (35)	-0.34	-1.0	-0.30	8.5x10^{-2NS}
L$^{1/2}$	ΔG$_{F-U}$	Alpha (33)	-0.54	0.87	-0.61	1.7x10^{-4}
L$^{1/2}$	ΔG$_{F-U}$	Beta (34)	-0.87	3.2	-0.69	5.2x10^{-6}
L$^{1/2}$	ΔG$_{F-U}$	Mixed (41)	-0.11	-4.5	-0.10	5.4x10^{-1NS}
L$^{1/2}$	ΔG$_{F-U}$	Maxwell (28)	-0.44	-1.4	-0.37	5.0x10^{-2NS}
log L	log k$_f$	Full (108)	-5.7	13	-0.69	2.0x10^{-16}
log L	log k$_f$	Two-State (73)	-4.9	12	-0.57	1.4x10^{-7}
log L	log k$_f$	Multi-State (35)	-8.0	18	-0.82	1.4x10^{-9}
log L	log k$_f$	Alpha (33)	-4.3	11	-0.75	4.3x10^{-7}
log L	log k$_f$	Beta (34)	-6.2	13	-0.74	5.5x10^{-7}
log L	log k$_f$	Mixed (41)	-5.7	13	-0.63	1.2x10^{-5}
log L	log k$_f$	Maxwell (28)	-0.69	3.6	-0.09	6.6x10^{-1NS}
log L	log k$_u$	Full (108)	-10	18	-0.78	6.5x10^{-23}
log L	log k$_u$	Two-State (73)	-11	20	-0.79	6.2x10^{-17}
log L	log k$_u$	Multi-State (35)	-11	21	-0.72	1.1x10^{-6}
log L	log k$_u$	Alpha (33)	-8.7	17	-0.81	9.4x10^{-9}
log L	log k$_u$	Beta (34)	-12	21	-0.87	2.1x10^{-11}
log L	log k$_u$	Mixed (41)	-6.8	11	-0.60	4.0x10^{-5}
log L	log k$_u$	Maxwell (28)	-5.4	8.6	-0.46	1.4x10^{-2}
log L	log k$_f$ = log k$_u$	Full (108)	-8.1	15	-0.76	1.4x10^{-21}
log L	log k$_f$ = log k$_u$	Two-State (73)	-9.3	17	-0.76	9.5x10^{-15}
log L	log k$_f$ = log k$_u$	Multi-State (35)	-7.6	15	-0.77	5.9x10^{-8}
log L	log k$_f$ = log k$_u$	Alpha (33)	-7.8	16	-0.83	3.1x10^{-9}
log L	log k$_f$ = log k$_u$	Beta (34)	-11	19	-0.87	3.2x10^{-11}

log L	log k_f = log k_u	Mixed (41)	-4.3	7.4	-0.59	$4.9\text{x}10^{-5}$
log L	log k_f = log k_u	Maxwell (28)	-3.2	5.6	-0.38	$4.8\text{x}10^{-2}$
log L	ΔG_{F-U}	Full (108)	-5.9	6.7	-0.53	$4.8\text{x}10^{-9}$
log L	ΔG_{F-U}	Two-State (73)	-8.3	11	-0.69	$1.4\text{x}10^{-11}$
log L	ΔG_{F-U}	Multi-State (35)	-4.6	4.6	-0.30	$7.6\text{x}10^{-2NS}$
log L	ΔG_{F-U}	Alpha (33)	-6.1	7.5	-0.65	$3.7\text{x}10^{-5}$
log L	ΔG_{F-U}	Beta (34)	-8.0	11	-0.70	$5.0\text{x}10^{-6}$
log L	ΔG_{F-U}	Mixed (41)	-1.5	-2.6	-0.11	$5.1\text{x}10^{-1NS}$
log L	ΔG_{F-U}	Maxwell (28)	-6.4	6.8	-0.43	$2.4\text{x}10^{-2}$

Individual correlations and linear fits are shown for subsets of the data as in Figure 4.1. Additionally, values for the commonly used dataset of Maxwell *et al.* [221] are shown for comparison

[a] Two-tailed probability value

[NS] Correlation is not significant at the 0.05 level ($5.0 \text{ x } 10^{-2}$)

$\Delta G_{F-U} = G_F - G_U = -RT \ln(k_f/k_u)$, where R is the gas constant and T is the absolute temperature in Kelvin, gives the Gibbs free energy of the folded state relative to the unfolded state

log k_f = log k_u where the thermodynamic stability of the protein is zero, is used to correct for the possible effects of differing stabilities

Relative Contact Order (RCO) is calculated as by Plaxco *et al.* [197]

Absolute Contact Order (ACO) is calculated as by Ivankov *et al.* [198]

Long Range Order (LRO) is calculated as by Gromiha *et al.* [210]

Short Range Order (SRO), local contact order (LocalCO), and non-local contact order (NonLocalCO) are calculated as by Zou and Ozkan [211]

Total Contact Distance (TCD) is calculated as by Zhou and Zhou [212]

Clustering Coefficient (CC) is calculated as by Micheletti [214]

Nα is calculated as by Ouyang and Liang [213]

In the case of all of the above parameters the original equations and cutoffs were used, except for definitions of contact, which were all made consistent with ACO by counting a contact between residues when any heavy atoms between the two residues were < 6.0 Å apart

Dependencies on length were suggested by various works [200, 251, 252, 199, 204, 203]

Table 4.3: Dataset

Protein Name	PDB (res)	Kin. State	Class	Len.	logk_f = logk_u	logk_f (s^{-1})^a	logk_u (s^{-1})^a	ΔG_{F-U} ($\frac{kcal}{mol}$)	Ref
Colicin E7 immunity protein	1AYI	Two	α	85	1.22	3.13	1.00	-2.90	[221]
Telomeric protein DNA-binding domain, human	1BA5	Two	α	49	0.69	2.56	0.52	-2.78	[225]
Immunoglobulin binding B-domain	1BDD (2-59)	Two	α	58	2.52	5.08	1.82	-4.44	[225]
16th domain of brain α-spectrin	1CUN (7-112)	Two	α	106	-0.87	2.08	-2.61	-6.40	[225]
17th domain of brain α-spectrin	1CUN (113-219)	Two	α	107	-1.48	1.48	-3.39	-6.64	[225]
FADD death-domain, human	1E41 (93-192)	Two	α	100	-0.26	2.95	-1.30	-5.81	[225]
Rap1 myb-domain, human	1FEX	Two	α	59	1.69	3.56	1.26	-3.14	[225]
Myb transforming protein	1IDY	Two	α	54	1.35	3.78	0.74	-4.15	[225]
Colicin E9 immunity protein	1IMQ	Two	α	85	-0.61	3.17	-0.83	-5.45	[221]
Trp-Cage Miniprotein	1L2Y	Two	α	20	5.65	5.43	4.99	-0.59	[225]
Lyme disease variable surface antigen	1L8W (29-335)	Two	α	307	-2.04	0.88	-3.68	-6.22	[221]
Lambda repressor	1LMB	Two	α	80	2.26	4.52	1.39	-4.27	[221]
Acyl-coenzyme A binding protein, cow	1NTI	Two	α	86	-0.13	3.04	-1.69	-6.46	[221]
Protein yjbJ	1RYK	Two	α	69	2.78	3.95	1.95	-2.73	[221]
BBA5 mini-protein	1T8J	Two	α	23	5.43	5.12	5.73	0.83	[225]
15th domain of brain alpha-spectrin	1U5P	Two	α	110	1.74	4.78	0.13	-6.34	[225]
Villin Headpiece	1VII	Two	α	36	4.60	4.08	2.30	-2.43	[248]

Dihydrolipolysine acetyltransferase, G. stearothermophilus	1W4G	Two	α	45	2.52	4.44	1.30	-4.28	[223]
Dihydrolipolysine succinyltransferase, E. coli	1W4H	Two	α	45	4.00	5.11	3.78	-1.82	[223]
Pyruvate dehydrogenase E2, P. aerophilum	1W4J	Two	α	51	3.26	5.32	2.74	-3.52	[223]
de novo designed 3-helix bundle	2A3D	Two	α	73	5.39	5.30	3.30	-2.73	[248]
Peripheral subunit-binding domain, Dihydrolipoamide acetyltransferase	2PDD	Two	α	42	4.26	4.26	2.35	-2.61	[250]
E3-binding domain of BBL	2WXC	Two	α	47	3.47	4.86	2.87	-2.73	[225]
Cold shock protein, B. caldolyticus	1C9O	Two	β	66	0.09	3.13	-0.17	-4.50	[225]
Cold shock protein, B. subtilus	1CSP	Two	β	67	1.17	2.82	1.00	-2.49	[225]
Formin Binding Protein 28	1E0L	Two	β	37	4.00	4.60	3.73	-1.18	[225]
WW prototype	1E0M	Two	β	37	3.34	3.87	3.08	-1.07	[225]
9th fibronectin domain	1FNF	Two	β	90	-0.96	-0.39	-1.26	-1.18	[225]
Cold shock protein, T. maritima	1G6P	Two	β	66	-1.13	2.74	-1.74	-6.10	[225]
Hisactophilin	1HCD	Two	β	117	-2.82	1.35	-4.69	-8.23	[249]
sso7d	1JIC	Two	β	62	0.26	3.04	-1.39	-6.04	[225]
Abp1 SH3 domain	1JO8	Two	β	58	-0.91	1.09	-1.17	-3.08	[221]
Fibronectin type III WL-12 chitinase A1	1K85 (559-644)	Two	β	86	-1.65	0.61	-3.04	-4.98	[225]
E2 component alpha-ketoacid dehydrogenase	1K8M	Two	β	87	-2.04	-0.39	-3.39	-4.09	[225]
Yes kinase-associated protein	1K9Q (5-44)	Two	β	40	3.04	3.65	2.91	-1.01	[225]
Internalin B SH3 domain	1M9S (391-466)	Two	β	76	0.04	1.74	-0.74	-3.38	[221]
Cold shock protein, E. coli	1MJC	Two	β	69	0.69	2.30	0.61	-2.31	[225]

Beta-hairpin	1PGB (41-56)	Two	β	16	5.21	5.21	5.21	0.00	[225]
PinWW	1PIN (6-39)	Two	β	34	3.95	4.04	1.82	-3.02	[225]
PI3 SH3 domain	1PNJ	Two	β	84	-2.08	-0.43	-3.17	-3.73	[225]
Oncoprotein p13mtcp1	1QTU (1-109)	Two	β	109	-3.69	0.00	-4.78	-6.52	[225]
Fyn SH3 domain	1SHF	Two	β	59	-0.87	2.13	-1.87	-5.45	[221]
Spectrin SH3 domain	1SHG	Two	β	57	-1.61	0.48	-2.08	-3.50	[221]
Src SH3 domain	1SRL	Two	β	56	0.17	1.91	-0.56	-3.38	[221]
Fibronectin type III tenascin	1TEN	Two	β	89	-1.91	0.48	-3.34	-5.21	[225]
18th module of muscle protein twitchin	1WIT	Two	β	93	-2.56	0.17	-3.56	-5.09	[225]
Sho1 SH3 domain	2VKN	Two	β	66	-0.35	0.92	-1.08	-2.73	[221]
Symfoil1	3O49	Two	β	123	-3.08	0.74	-3.91	-6.34	[79]
Symfoil4P	3O4D	Two	β	123	-4.69	2.13	-5.99	-11.08	[79]
ThreeFoil	3PG0	Two	β	140	-7.43	-3.87	-8.95	-6.93	[70]
Muscle acylphosphatase	1APS	Two	αβ	98	-3.17	-0.69	-3.91	-4.38	[221]
C-terminal domain of protein L9	1DIV (58-149)	Two	αβ	92	-1.91	1.43	-3.43	-6.64	[221]
N-terminal domain of protein L9	1DIV (1-56)	Two	αβ	56	0.83	2.87	0.00	-3.90	[221]
LysM Domain	1E0G	Two	αβ	48	1.61	3.04	1.02	-2.75	[225]
FK506 binding protein	1FKB	Two	αβ	107	-2.26	0.69	-3.52	-5.75	[221]
Apoflavodoxin, Anabaena	1FTG	Two	αβ	168	-0.09	1.22	-1.17	-3.26	[225]
Tm1083	1J5U	Two	αβ	123	-0.39	2.97	-2.28	-7.17	[221]
Chemotaxis protein CheW	1K0S	Two	αβ	143	-2.17	3.21	-5.25	-11.55	[221]
Cyclophilin A	1LOP	Two	αβ	164	-1.30	2.87	-4.52	-10.07	[225]
Ribosomal protein L23	1N88	Two	αβ	96	-1.04	0.87	-1.69	-3.50	[221]
ADAh2	1O6X	Two	$\alpha\beta$	81	0.65	2.95	-0.17	-4.27	[221]
B1 domain of protein G	1PGB	Two	$\alpha\beta$	56	-0.17	2.74	-0.74	-4.74	[221]

Histidine containing phosphocarrier protein	1POH	Two	$\alpha\beta$	85	-1.30	1.17	-2.69	-5.27	[225]
C-terminal domain of spore coat protein S	1PRS (91-173)	Two	$\alpha\beta$	83	-3.52	-0.87	-4.04	-4.32	[225]
N-terminal domain spore coat protein S	1PRS (1-90)	Two	$\alpha\beta$	90	-2.00	1.30	-4.08	-7.35	[225]
Ras binding domain	1RFA	Two	$\alpha\beta$	78	-0.09	3.65	-1.22	-6.64	[221]
Ribosomal protein S6	1RIS	Two	$\alpha\beta$	97	-1.69	2.65	-3.60	-8.53	[221]
Src SH2 domain	1SPR	Two	$\alpha\beta$	103	-0.61	3.78	-1.52	-7.23	[221]
Ubiquitin	1UBQ	Two	$\alpha\beta$	76	-0.65	3.17	-2.95	-8.35	[221]
Spliceosomal protein U1A	1URN	Two	$\alpha\beta$	96	-0.17	2.01	-5.09	-9.68	[221]
Common-type acylphosphatase	2ACY	Two	$\alpha\beta$	98	-1.91	0.35	-2.82	-4.32	[225]
Chymotrypsin inhibitor 2	2CI2	Two	$\alpha\beta$	64	-1.52	2.52	-4.47	-9.54	[221]
B1 domain of Protein L	2PTL (18-77)	Two	$\alpha\beta$	60	-0.61	1.78	-1.43	-4.38	[221]
Pit1 homeodomain	1AU7 (103-160)	Multi	α	58	3.21	4.21	2.08	-2.90	[225]
FKBP-Rapamycin binding domain	1AUE	Multi	α	92	-1.04	2.61	-2.35	-6.75	[225]
p19ink4d CDK inhibitor	1BD8	Multi	α	156	-0.83	1.26	-1.00	-3.08	[225]
Engrailed homeodomain	1ENH	Multi	α	54	3.52	4.56	3.30	-1.72	[225]
Acyl-coenzyme A binding protein, yeast	1ST7	Multi	α	86	-0.17	3.69	2.78	-1.24	[225]
FF domain, human hypa	1UZC	Multi	α	69	1.00	3.30	0.61	-3.67	[225]
Tumour suppressor protein p16	2A5E (9-156)	Multi	α	148	0.17	1.52	0.09	-1.95	[225]
Phage 434 cro protein	2CRO	Multi	α	64	0.13	1.61	-0.22	-2.49	[225]
T4 lysozyme	2LZM	Multi	α	164	-2.48	1.78	-6.08	-10.72	[225]
Myotrophin	2MYO	Multi	α	118	-0.13	2.04	-1.39	-4.68	[225]
Ileal lipid binding protein	1EAL	Multi	β	127	-1.22	0.56	-2.13	-3.67	[225]
Intestinal fatty acid binding protein	1IFC	Multi	β	131	-1.22	1.48	-2.00	-4.74	[225]

FGF-1	1JQZ	Multi	β	136	-2.87	0.56	-3.08	-4.98	[79]
Cellular retinol binding protein II	1OPA	Multi	β	133	-2.17	0.61	-4.00	-6.28	[225]
Barnase	1RNB	Multi	β	109	-1.87	1.13	-3.95	-6.93	[225]
Titin IG repeat 27	1TIU	Multi	β	89	-3.00	1.56	-3.30	-6.64	[225]
10th type III fibronectin domain	1TTF	Multi	β	94	-0.22	2.39	-3.65	-8.23	[225]
Apical domain of GroEL	1AON (191-345)	Multi	$\alpha\beta$	155	-1.48	-0.65	-2.48	-2.49	[225]
Third PDZ domain from PSD-95	1BFE	Multi	$\alpha\beta$	110	-0.09	1.30	-1.48	-3.79	[225]
Barstar	1BTA	Multi	$\alpha\beta$	89	-0.61	1.48	-1.17	-3.61	[225]
Cellular retinoic acid binding protein I	1CBI	Multi	$\alpha\beta$	136	-2.91	-1.39	-4.26	-3.91	[225]
PDZ2 domain from PTP-BL	1GM1	Multi	$\alpha\beta$	94	-0.52	0.35	-1.00	-1.84	[225]
Hydrogenase maturation protein	1GXT	Multi	$\alpha\beta$	88	-0.35	1.91	-2.95	-6.64	[225]
Indole-3-glycerolphosphate synthase	1IGS (27-248)	Multi	$\alpha\beta$	222	-3.87	-2.00	-5.91	-5.33	[225]
Staphylococcal nuclease	1JOO	Multi	$\alpha\beta$	149	-2.30	1.00	-3.65	-6.34	[225]
C-terminal domain of phosphoglycerate kinase	1PHP (176-394)	Multi	$\alpha\beta$	219	-2.21	-1.69	-4.69	-4.09	[225]
N-terminal domain of phosphoglycerate kinase	1PHP (1-175)	Multi	$\alpha\beta$	175	-0.56	1.00	-1.82	-3.85	[225]
Trp-Synthase α-subunit	1QOP (Chain A)	Multi	$\alpha\beta$	265	-2.48	-1.09	-3.87	-3.79	[225]
Dihydrofolate reductase	1RA9	Multi	$\alpha\beta$	159	-2.26	-1.39	-2.65	-1.72	[225]
Cell-cycle regulatory protein p13suc1	1SCE	Multi	$\alpha\beta$	97	-0.43	1.82	-2.65	-6.10	[225]
Carbonic anhydrase	1V9E	Multi	$\alpha\beta$	259	-4.60	-1.82	-10.42	-11.73	[225]

Ribonuclease H1, E. coli	2RN2	Multi	$\alpha\beta$	155	-2.00	0.04	-5.21	-7.17	[225]
Villin 14T	2VIK	Multi	$\alpha\beta$	126	-0.69	2.17	-1.78	-5.39	[225]
Chemotactic protein	3CHY	Multi	$\alpha\beta$	128	-0.56	0.43	-1.91	-3.20	[225]
Ribonuclease H1, C. tepidum	3H08	Multi	$\alpha\beta$	139	-1.74	0.83	-6.08	-9.42	[225]
Average (# proteins)									
Full (108)					-0.28	2.00	-1.50	-4.77	
Two-State (74)					0.02	2.42	-1.03	-4.71	
Multi-State (34)					-1.06	1.08	-2.50	-4.89	
Alpha (33)					1.54	3.50	0.68	-3.84	
Beta (34)					-0.83	1.54	-1.78	-4.53	
Mixed (41)^b					-1.30	1.17	-3.01	-5.71	
Maxwell (28)^c					-0.54	2.27	-1.81	-5.56	
Min (# proteins)									
Full (108)					-7.43	-3.87	-10.42	-11.73	
Two-State (74)					-7.43	-3.87	-8.95	-11.55	
Multi-State (34)					-4.60	-2.00	-10.42	-11.73	
Alpha (33)					-2.48	0.88	-6.08	-10.72	
Beta (34)					-7.43	-3.87	-8.95	-11.08	
Mixed (41)^b					-4.06	-2.00	-10.42	-11.73	
Maxwell (28)^c					-3.17	-0.69	-5.23	-11.55	
Max (# proteins)									
Full (108)					5.65	5.43	5.73	0.83	
Two-State (74)					5.65	5.43	5.73	0.83	
Multi-State (34)					3.52	4.56	3.30	-1.24	
Alpha (33)					5.65	5.43	5.73	0.83	
Beta (34)					5.21	5.21	5.21	0.00	
Mixed (41)^b					1.61	3.78	1.02	-1.72	

Maxwell (28)[c]	2.78	4.52	1.95	-2.73
Standard Deviation (# proteins)				
Full (108)	2.40	1.88	2.92	2.53
Two-State (74)	2.60	1.83	2.92	2.52
Multi-State (34)	1.66	1.66	2.69	2.58
Alpha (33)	2.27	1.36	2.59	2.25
Beta (34)	2.59	1.75	2.90	2.42
Mixed (41)[b]	1.29	1.63	2.05	2.57
Maxwell (28)[c]	1.32	1.23	1.85	2.37

[a] all values refer to 25 °C

[b] Mixed refers to the $\alpha\beta$ structure class

[c] the set "Maxwell" is listed as ref [221]

log k_f = log k_u is the experimentally determined mid-transition (denaturation midpoint) folding (equal to unfolding) rate, where the thermodynamic stability of the protein is zero

ΔG_{F-U} = G_F - G_U = -RT ln(k_f/k_u), where R is the gas constant and T is the absolute temperature in Kelvin (298.15), gives the Gibbs free energy of the folded state relative to the unfolded state

Table 4.4: Single mutation effects on folding and unfolding rates

Protein	Number of Mutations	Average change in k_f (x-fold)^a	Average change in k_u (y-fold)^a	Relative change (y-fold / x-fold)
Muscle AcP	22	3.16	8.12	2.57
FKBP12	34	2.25	29.64	13.18
L23	17	2.03	104.11	51.35
CTL9	24	3.98	131.22	33.00
α-spectrin SH3 (D48G)	14	5.12	13.45	2.63
CI2	65	3.32	65.21	19.64
Src-SH3	54	2.79	9.22	3.30
Protein L	68	1.75	44.83	25.60
Fyn-SH3	34	7.01	26.90	3.84
bACBP	30	3.54	134.25	37.89
Ubiquitin	27	4.41	232.79	52.75
Protein G	31	2.87	14.97	5.22
ADA2h	18	1.78	4.70	2.64
NTL9	24	1.94	18.48	9.54
sso7d	20	2.15	7.19	3.35
CspB	21	3.01	7.89	2.62
Im9	25	12.32	83.70	6.80
cyt b562	39	2.82	173.39	61.59
yACBP	18	5.28	161.14	30.54
FBP28 WW	45	1.40	2.83	2.02
E3DB (F166W)	22	3.30	16.80	5.08
BdpA (Y15W)	45	2.84	12.97	4.57
BdpA (N29H,Q33W)	20	1.70	9.32	5.49
BpdA (E48W)	20	2.35	6.39	2.72
POB (L146A, Y166W)	22	3.37	6.99	2.07
Per Protein Average^b	25	3.46	53.06	**15.34**

Per Mutation Average[c]	759	3.30	52.58	**15.95**

[a] change is calculated as eˆ|ln k_{wt} - ln k_{mut}|. Thus, a two-fold slower rate or two-fold faster rate for the mutant would both be a 2.0 fold change in rate. Reported values are the average across all mutations for that protein

[b] average across the dataset weighting each protein equally

[c] average across the dataset weighting each mutation equally

It should be noted that the vast majority of mutations in the dataset [204] are simple truncation mutations and hence, the relationships shown may not be representative of a more general set of mutations

Chapter 5

Topological Complexity and Kinetic Stability

5.1 Context

This chapter is a pre-print version of “Designed protein reveals structural determinants of extreme kinetic stability”, published in the Proceedings of the National Academy of Sciences of the United States of America in 2015, of which I am the first author [70]. While the previous chapter demonstrated the connection between high topological complexity and slow unfolding rates, in this chapter the connection to kinetic stability and its associated benefits is examined. In particular, slow unfolding — when in the context of two-state unfolding kinetics — results in a protein that is effectively trapped in the native state, regardless of its thermodynamic stability. This can allow proteins to survive particularly harsh solvent conditions, the presence of protease, and high temperatures.

I determined the experimental kinetics of ThreeFoil and its resistance properties in collaboration with S. Martha Ma. I determined the resistance of a number of additional proteins in collaboration with Hitesh Rafalia. I collected data for numerous resistant proteins in collaboration with Ke Xia. I performed overall analysis of various datasets as well as collecting and analyzing all data found in the supplemental information (not including MassSpec and Diagonal 2D SDS-PAGE which were performed by Ke Xia, and not including *ab initio* folding of OneFoil which was done by Kyle Trainor). Shachi Gosavi performed and analyzed all Gō-modelling simulations. I wrote the manuscript along with Shachi Gosavi and Elizabeth M. Meiering.

5.2 Summary

The design of stable, functional proteins is difficult. Improved design requires a deeper knowledge of the molecular basis for design outcomes and properties. We previously used a bioinformatics and energy function method to design a symmetric superfold protein composed of repeating structural elements with multivalent carbohydrate-binding function, called ThreeFoil. This and similar methods have produced a notably high yield of stable proteins. Using a battery of experimental and computational analyses we show that in spite of its small size and lack of disulfide bonds, ThreeFoil has remarkably high kinetic stability and its folding is specifically chaperoned by carbohydrate binding. It is also extremely stable against thermal and chemical denaturation and proteolytic degradation. We demonstrate that the kinetic stability can be predicted and modelled using absolute contact order (ACO) and long-range order (LRO), as well as coarse-grained simulations; the stability arises from a topology that includes many long-range contacts which create a large and highly cooperative energy barrier for unfolding and folding. Extensive data from proteomic screens and other experiments reveal that a high ACO/LRO is a general feature of proteins with strong resistances to denaturation and degradation. These results provide new and tractable approaches for predicting resistance and designing proteins with sufficient topological complexity and long-range interactions to accommodate destabilizing functional features as well as withstand chemical and proteolytic challenge (5.1).

5.3 Significance

Much research has focused on the molecular basis for protein thermodynamic stability; by comparison, kinetic stability is much less understood. Thermodynamics define the equilibrium fraction of unfolded protein while kinetics define the rate of unfolding; the latter can be of great practical importance for ensuring a protein remains folded under biological conditions. Using extensive experimental and modelling analyses we show that ThreeFoil, a small glycan binding protein without disulfides, exhibits outstanding kinetic stability against chemical denaturation and proteolytic degradation. We demonstrate that high kinetic stability is successfully modelled in terms of extensive long-range intramolecular interactions. These results show that topological complexity is a key determinant of kinetic stability which should help in designing proteins to withstand harsh conditions (5.1).

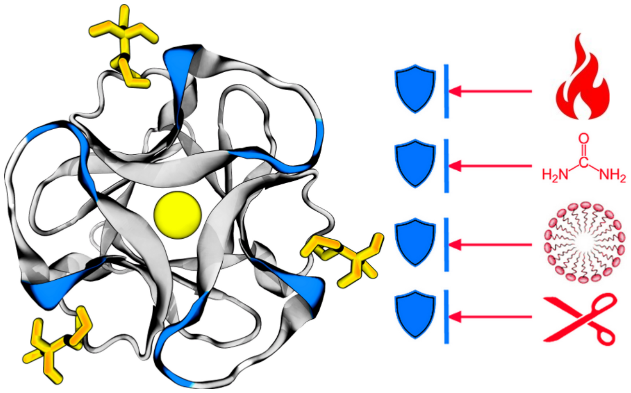


Figure 5.1: Topological complexity promotes survival under harsh conditions. Owing to long and well-structured loops (blue), which make extensive long-range contacts with residues distant in sequence, ThreeFoil and other topologically complex proteins are able to resist degradation even in harsh environment conditions. Such kinetically stable proteins can resist: high temperature, denaturants (such as urea), detergents, and proteases (shown in read from top to bottom).

5.4 Introduction

The design of proteins with a desired stable fold and function is a much sought after goal. While impressive recent successes have been reported in designing both natural and novel protein functions and/or structures [44, 253, 39, 254, 19, 21], design remains difficult, often requiring multiple rounds of iterative improvements [79, 14, 15, 255]. In depth biophysical characterization of protein design outcomes and an understanding of their molecular basis have been limited, and these are critical for improving future designs. Combining designed function with structure is particularly difficult, in part because functional sites tend to be sources of thermodynamic instability [154, 256] and folding frustration [257, 258, 259]. We investigate how an approach that considers both structure and function from the outset may be used to overcome such obstacles. Furthermore, we demonstrate how kinetic and related stabilities against denaturation can be rationally designed.

A promising emerging paradigm for protein design is the repetition of modular structural elements [44, 253, 19, 21, 79, 258, 74, 45, 47, 64]. This approach can simplify the design process and build on aspects of the evolution of natural repetition in proteins, as well as incorporate the inherent multivalent binding functionality of such structures [44, 48]. Internal structural symmetry, resulting from the repetition of smaller elements of structure, is very common in natural proteins, with ~20% of all protein folds [43] and the majority of the most populated globular protein folds (superfolds) [48] containing internal structural symmetry. Recent design successes, for helical proteins [19, 21], repeat proteins [47, 64, 67] and symmetric superfolds [44, 253, 79, 74, 45, 50, 161, 75, 52] recommend the simplification of the design process by using repetitive/symmetric folds as a particularly effective strategy.

The β-trefoil superfold is an interesting test case for design by repetition as bioinformatics analysis has revealed multiple and recent instances of the evolution of distinct proteins with this symmetric fold [44]. The fold consists of three repeats, each containing four β-strands, and is adopted by numerous superfamilies with highly diverse binding functions [122]. Our design of a completely symmetric β-trefoil, ThreeFoil (Figure 5.2), used a hypothetical multivalent carbohydrate binding template and mutated 40 of the 141 residues [44]. The mutations were based on a combination of consensus design using a limited set of close homologs (in order to preserve function), and energy scoring using Rosetta [260]. The design was successful on the first attempt, producing a soluble, well-folded, and functional monomer with very high resistance to structural fluctuations as indicated by high resistance to thermal denaturation and limited amide H/D exchange [44].

Here, we use a battery of biophysical and computational methods to perform an in depth analysis of Threefoil, which shows that it has remarkably slow unfolding and folding kinetics compared to natural and designed proteins due to an unusually high transition state energy barrier. Such kinetic stability against unfolding has been studied little to date. Furthermore, Threefoil is extremely resistant to chemical denaturation and proteolytic degradation. Analyses using Absolute Contact Order (ACO) [198] and Long-Range Order (LRO) [210] as well as Gō model folding simulations [261, 226, 262] show that ThreeFoil's resistance can be explained by the high cooperativity of its folded structure, which includes many long-range interactions. Simulations also show that non-native interactions or folding frustration arising from protein symmetry [263] do not create long-lived traps during folding or account for the high barrier. They also explain how ligand binding can chaperone folding, which can be an added advantage of designing the fold and function together. Notably, additional analyses using whole proteome screening and other experiments show that proteins with similar resistances as ThreeFoil generally have high ACO/LRO values. Thus, the design method used for ThreeFoil and the strategy of designing folds with many long-range contacts may be useful for designing functional proteins with high resistance to denaturation and degradation, as may be needed for challenging biotechnology applications.

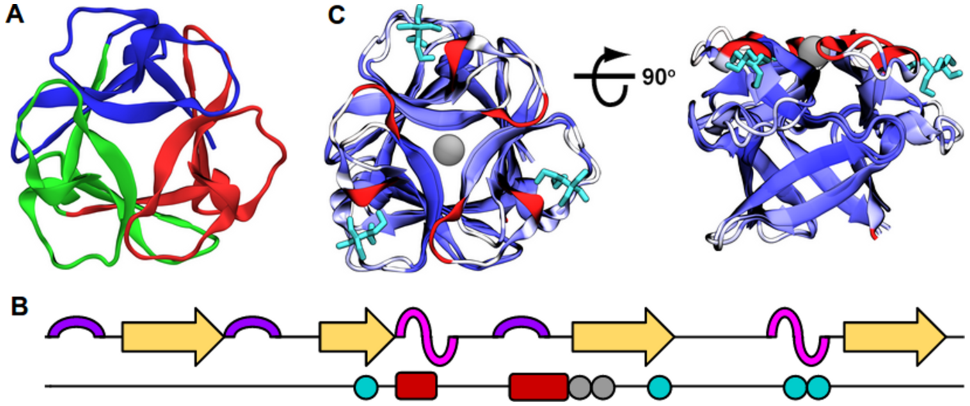


Figure 5.2: Design of ThreeFoil. (A) ThreeFoil (PDB: 3PG0) illustrating its three identical peptide subdomains (red, green, blue). (B) ThreeFoil's secondary structure: turn (purple), β-strand/bridge (yellow), and 3/10-helix (magenta) and ligand binding residues indicated by colored circles and insertions shown in red. (C) Comparison of ThreeFoil with the independently designed Symfoil (PDB: 3O4D, 15% sequence identity), shown along (left) and across (right) the axis of symmetry. Backbones are colored by RMSD between the two structures (blue to white, 0 to 5 Å), with insertions in the loops of ThreeFoil relative to Symfoil colored red. ThreeFoil's bound sodium shown in grey and bis-tris, which binds in the conserved sugar binding sites, shown in cyan.

5.5 Results

5.5.1 ThreeFoil has extremely slow kinetics and substantial thermodynamic stability

To better understand the basis for ThreeFoil's very high apparent melting temperature (>90 °C) and slow amide exchange [44], we measured its folding kinetics and thermodynamic stability using chemical denaturation. ThreeFoil is extremely resistant to chemical denaturation, remaining folded in high concentrations of urea, with unfolding only observable after very long incubation times in high concentrations of the stronger denaturants guanidinium chloride and guanidinium thiocyanate (GuSCN) (Figure 5.3, Figure 5.7). ThreeFoil's half-life for unfolding in the absence of denaturant is ∼8 years, while its folding half-life is on the order of an hour (Table 5.1). A comparison against natural and designed proteins of varying structural classes and lengths illustrates how unusually slow these kinetics are (Figure 5.4). Despite ThreeFoil's slow kinetics, unfolding is highly reversible and the rate constants measured by multiple optical probes vary linearly with denaturant concentration, indicating a 2-state transition between folded and unfolded states

(Figure 5.3A, Figure 5.7). The very slow kinetics are indicative of a high free energy (un)folding transition state (Figure 5.3B). Similarly, a high transition state barrier underlies the extremely slow unfolding of α-lytic protease [264]; however, unlike this prototypical kinetically stable protein which is thermodynamically unstable, ThreeFoil possesses substantial thermodynamic stability of ∼6 kcal/mol (Table 5.1).

Various studies have found evidence that repetition in proteins can slow kinetics by creating folding frustration [263, 265]. To further examine the role of sequence repetition on kinetics, we compare ThreeFoil with another fully symmetric β-trefoil, Symfoil, which was obtained using iterative rounds of rational design and sequence selection [79]. Symfoil has <15% sequence identity to ThreeFoil and, though it has a higher thermodynamic stability of ∼11 kcal/mol, it both folds and unfolds much faster (1 million-, and 400-fold, respectively, Table 5.1, Figure 5.4). Thus, symmetry does not *a priori* result in kinetic stability. Despite having a nearly identical core secondary structure to Symfoil, ThreeFoil has additional length and interactions for a set of loops involved in its carbohydrate binding function (Figure 5.2B,C). By contrast, heparin binding function, including binding residues in a loop of Symfoil's template protein, acidic fibroblast growth factor, were eliminated during the many iterations of the design process [79]. As ThreeFoil's longer loops surround and create its ligand binding sites, we investigated the structure of these sites during folding.

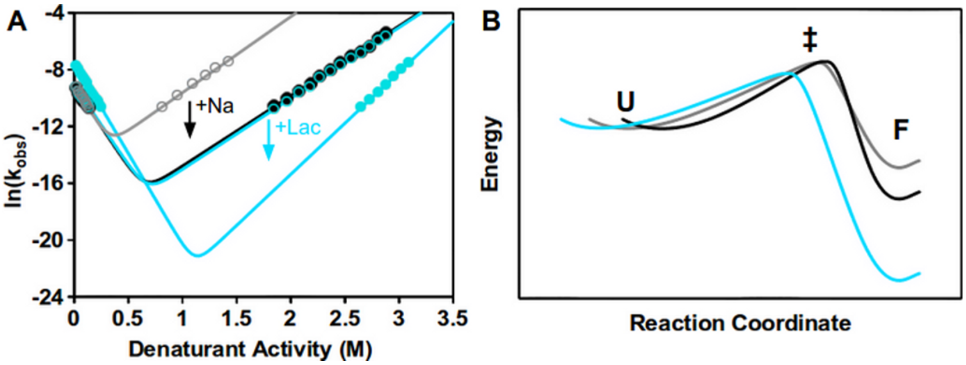


Figure 5.3: Folding and unfolding kinetics of ThreeFoil are modulated by ligand binding. (A) Chevron plots of observed folding and unfolding rate constants (in s^{-1}) in GuSCN were determined by fluorescence. Measurements were without sodium (grey open circles), with sodium (300 mM, black filled circles) or sodium and 50 mM of either lactose (cyan filled circles) or sucrose (cyan open circles). (B) Energy diagram corresponding to the kinetic measurements (coloring as in A). The energy axis is given by $-RT\ln(k_{obs})$ and the reaction coordinate follows the change in solvent accessible surface area as measured by m_f and m_u. The folded (F), transition (‡) and unfolded (U) states are indicated. Unfolded state energies and folded state reaction coordinates are set equal to facilitate comparisons.

5.5.2 Formation of ligand binding sites during folding

Measurements of the kinetics of folding/unfolding in the presence of ligands can be used to monitor the formation of ligand binding sites [266]. ThreeFoil has a single binding site for sodium, which is coordinated by three sets of residues distant in sequence, and three carbohydrate binding sites, which have binding residues close in sequence (Figure 5.2B,C) [44]. The carbohydrate sites bind lactose and are highly specific for glycans with terminal galactose in a β-1,4 linkage (Figures 5.8A and 5.9). Sodium decreases the unfolding rate but has no effect on the folding rate, thus it binds only to the folded state and not to the transition state (Figure 5.3B). In contrast, lactose not only decreases the unfolding rate but also increases the folding rate, indicating partial formation of the lactose binding sites in the transition state ensemble. The kinetic effects of lactose are specific and not general solvent properties, as no kinetic changes are observed for sucrose (Figure 5.3, Table 5.1).

Interestingly, the addition of lactose also increases the denaturant-dependence of stability, m, which reports on the extent of solvent accessible surface burial for a structural transition. The m increases from a value that is 68% of that expected for a protein of this size to 85% (Table 5.1). An increase in m may arise from increased burial of hydrophobic residues in the folded protein and/or decreased residual structure in the unfolded protein. Multiple lines of evidence support the latter explanation. Circular dichroism (CD) and NMR (Figure 5.9 and Figure 5.10, respectively) experiments for folded ThreeFoil show no significant change in native structure upon lactose binding. Also, anilinonapthalenesulfonic acid (ANS), a dye that binds clusters of exposed hydrophobic groups, shows no binding to folded or denatured Threefoil (Figure 5.11). There is no apparent change in CD upon adding lactose to denatured ThreeFoil (Figure 5.8B); however, the CD spectra of denatured ThreeFoil show evidence for non-random structure. Similarly, quantitative CD analysis for OneFoil, a peptide consisting of just one of the constituent repeats of ThreeFoil, shows it has ∼half of the β-structure observed in folded ThreeFoil (Figure 5.8C). OneFoil shows no evidence for stable structure formation by NMR though [44]. These experiments strongly suggest the presence of fluctuating residual structure in the ensemble of denatured conformers for ThreeFoil. Together with folding simulations (described below), the results indicate that lactose binding enhances folding not only by binding to the transition state, but also by binding weakly to some conformations in the denatured ensemble and so decreasing non-native residual structure.

Other studies have also shown that different types of ligands (e.g. metals, heme, nucleotides) can bind to partially folded proteins (denatured, intermediate, and transition states) and so promote folding [267, 268, 269]. Thus, ligands may not only stabilize the native state, but also promote and chaperone protein folding by binding to other states and thereby smooth the folding energy landscape. In this way, ligand binding can increase the foldability of the protein when structure and function are designed concurrently rather than separately.

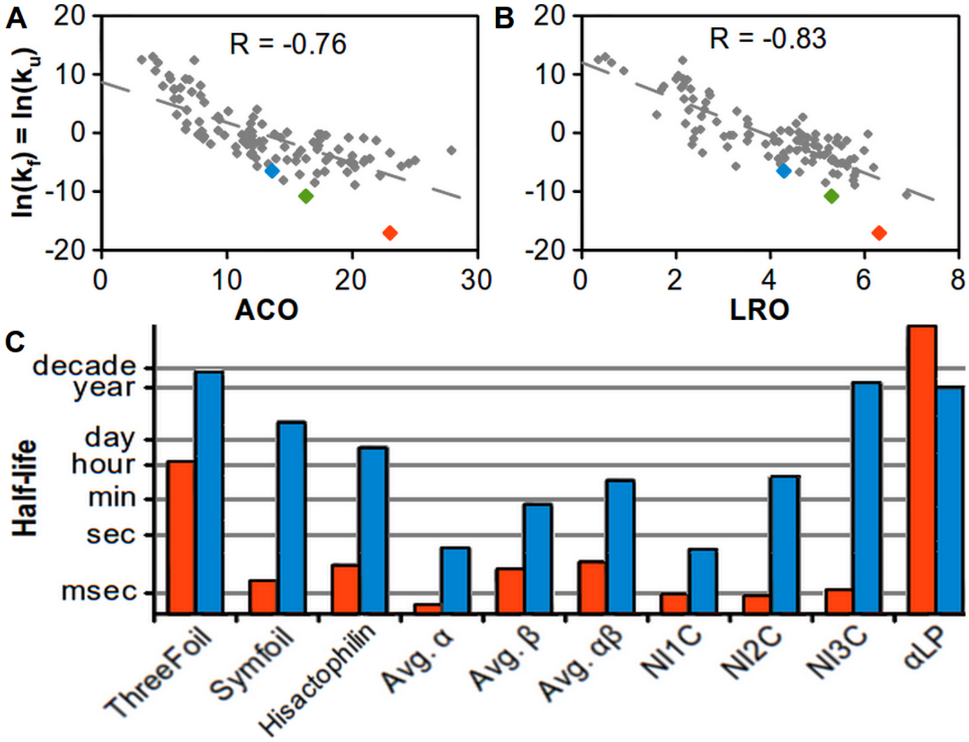


Figure 5.4: ThreeFoil folding/unfolding kinetics are extremely slow compared to other proteins. Rate constants (grey diamonds) at the transition mid-point ($\ln(k_f) = \ln(k_u)$) for a large dataset of proteins (Table 5.2) [78], are correlated with (A) ACO and (B) LRO. β-trefoil proteins: ThreeFoil (orange), Symfoil (green), and Hisactophilin (blue) are highlighted. (C) The half-lives for folding (orange) and unfolding (blue) are shown for β-trefoils and the averages for the large dataset in each major structural class (α, β, αβ). The prototypical kinetically stable protein α-lytic protease is shown for comparison [264]. Ankyrin proteins with 1-3 consensus designed internal repeats (NI1C to NI3C) illustrate the effect of increasing interface area and cooperativity [67].

5.5.3 Modelling reveals the molecular mechanisms of ThreeFoil's ligand binding and slow kinetics

The ligand binding loops make extensive contacts with distant residues in the primary sequence and so increase the ACO and LRO for ThreeFoil, which are notably high (Figure 5.4A,B). ACO/LRO are measures of topological complexity based on the sequence separation of contacting residues in the folded protein. We have shown recently that the rates of protein folding and unfolding both decrease with increasing ACO/LRO [78]. LRO provides a more linear and stronger correlation and is normalized for increasing protein size, which also slows (un)folding [78, 270]. As ACO/LRO provide just a simple measure of protein structural complexity, we used Gō models to further define the molecular origins of ThreeFoil's high barrier.

Gō models encode the structure of the folded protein in their energy functions [261, 226, 262] and can be used to understand at higher resolution the effects of protein topological complexity on folding. Molecular dynamics (MD) simulations of such models for diverse proteins can capture trends in barrier heights as well as mechanistic details of folding [226]. Here, a coarse-grained Gō model shows that ThreeFoil has a particularly high free energy barrier (Figure 5.5A). Also, in the structure of the transition state ensemble (Figure 5.5B,C) residues around the carbohydrate binding site of the second repeat are almost completely folded. Therefore, lactose may bind to and lower the energy of the transition state ensemble and so increase the folding kinetics, while unfolding kinetics are slowed owing to even stronger lactose binding to the folded state (Figure 5.3). In contrast, the residues in the sodium binding site are quite unstructured early in the transition state (Figure 5.5B,C) showing that sodium only binds the folded state and therefore slows unfolding with no effect on folding. Thus, the Gō model simulations rationalize ThreeFoil's slow experimental kinetics and also provide a molecular explanation for the kinetic effects of ligands. In addition, while a simple calculation of ACO/LRO indicates that ThreeFoil should un/fold slower than Symfoil and Hisactophilin (see below) (Figure 5.4A,B), the more detailed Gō model simulations better capture the variations in these rates (Figures 5.5G, 5.12).

The largest differences between Threefoil and Symfoil are in the β2-β3 loops, at the edge of ThreeFoil's carbohydrate binding sites (Figure 5.2B,C). Consequently, the differences in the contact maps of the two also occur mostly in the contacts of the β2-β3 loops, with Symfoil having shorter loops and fewer contacts in this region. In order to understand whether the high barrier for ThreeFoil is caused by differences in the length, conformation and packing of the β2-β3 loops or by differences in packing for the rest of the protein a hybrid construct (HYB) was created that has the Symfoil backbone with the ThreeFoil contact map; this construct has almost the same barrier as Symfoil (SymF) (Figure 5.5G). This indicates that the differences responsible for the higher barrier are the β2-β3 loops. To further define how the conformation and packing of the β2-β3 loops increases the barrier, a mutant of ThreeFoil (MUT1) was created with all long-range interactions of the β2-β3 loops deleted (Figure 5.5D,E). The mutation lowered the barrier height of MUT1 leaving it similar to that of HYB and SymF (Figure 5.5G). These results show that the packing of the β2-β3 loops of a given repeat with parts of the other repeats create long-range contacts that markedly increase the barrier height. In order to test the effect of other long-range contacts (which reduce the overall ACO by an equivalent amount), a control mutant (MUT2) was created where the same number of other contacts with similar sequence separations were deleted (Figure 5.5D,F). The free energy barrier of MUT2 is similar to that of both MUT1 and HYB. Thus, the kinetic stability of a protein can be reduced by either a large loss in packing density localized in the structure (as in MUT1) or by an additive effect from many losses across the structure (as in MUT2). To further confirm the correlation between ACO/LRO, barrier heights and kinetic stability, we also simulated Hisactophilin, a natural β-trefoil with a low barrier (Table 5.1). As expected, the low ACO/LRO Hisactophilin has a much lower folding free energy barrier (Figure 5.5G; green profile) than that of either ThreeFoil or SymFoil and has the lowest kinetic stability (Figure 5.4). Distinct functional features for Hisactophilin, namely a "hole" within its hydrophobic core, contribute to its low ACO/LRO and barrier [236].

In principle, the internal symmetry of ThreeFoil might also contribute to slow folding by creating misfolded intermediates arising from internal subdomain swapping, analogous to domain swapping observed or inferred for proteins containing longer stretches of repeated sequence [263, 265]. Such trapping was tested as a cause of ThreeFoil's slow kinetics using simulations performed with the addition of symmetric non-native contacts between the repeats. The results indicate that close to the transition midpoint non-native interactions arising from symmetry do not create significant trapping (Figure 5.13). Thus, compared to the longer proteins the shorter repeat length of ThreeFoil aided by local structure formation, likely limits slowing of folding due to non-native inter-repeat interactions.

MD simulations were also used to investigate the effect of non-native residual structure (in the unfolded ensemble) on ThreeFoil folding. We performed simulations where the local structure of all three repeats was biased to both the native ThreeFoil structure (as above) and to the most common OneFoil structure obtained in Rosetta *ab initio* simulations. The tertiary contacts between the repeats were calculated only from the native ThreeFoil structure. The ThreeFoil tertiary contacts appear to suppress the intra-OneFoil non-native interactions and these non-native interactions do not create significant trapping (Figure 5.14). Lastly, simulations of just OneFoil (including both native and non-native structural biases) confirm that the presence of ligand, modelled as a strengthening of intra-binding-residue contacts, greatly suppresses the formation of non-native structure (Figure 5.14A,B). Overall, the simulations explain the effects of ligands during folding while also revealing that the extreme kinetic stability of ThreeFoil arises from its native topology and is unlikely to be caused by non-native traps on the folding free energy landscape.

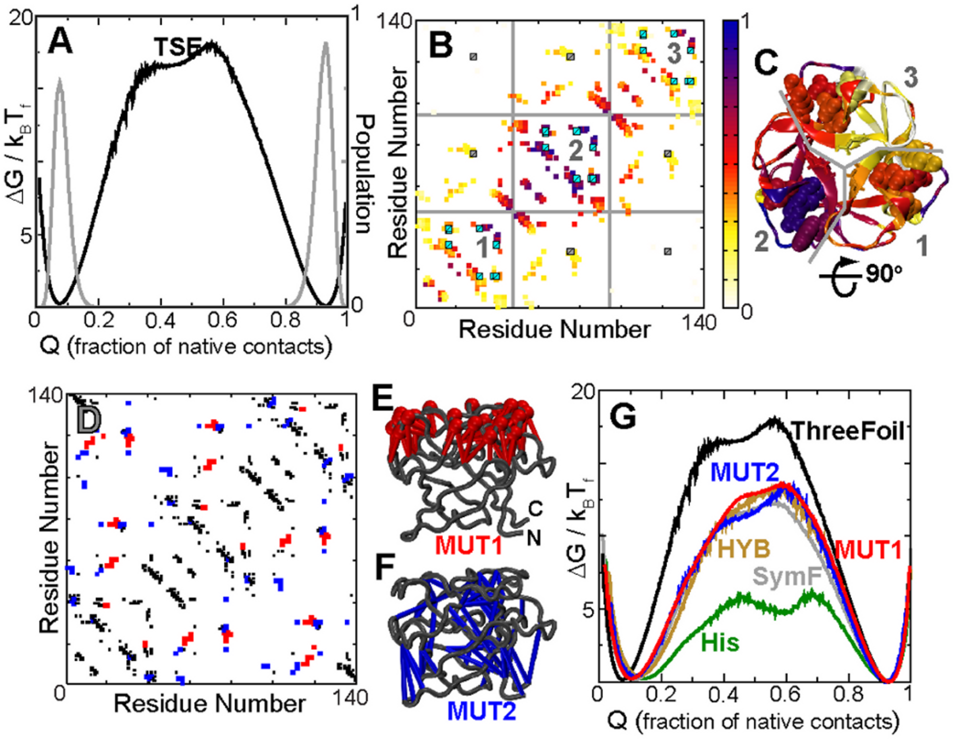


Figure 5.5: Structure-based simulations reveal the molecular origins of ThreeFoil's large kinetic barrier. (A) The folding free energy of ThreeFoil in units of $k_B T_f$ (left Y axis) is plotted at the transition midpoint (T_f) as a function of the fraction of native contacts (Q) in black. The population distribution is plotted in grey (right Y axis). The protein populates only the unfolded state (Q ∼0.1) and the folded state (Q ∼0.9). The two curves were calculated from simulations of a ThreeFoil model using all contacts shown in D. (B) Contact map of the transition state ensemble (TSE, Q ∼0.35 in A), colored based on degree of structure, with 1 indicating native levels of structure and 0 no structure. Contacts between lactose binding site residues (cyan) and sodium binding residues (grey) are shown. The numbered squares contain intra-trefoil contacts (see C). (C) Average level of structure derived from B (same coloring) illustrated for ThreeFoil partitioned into its three repeats by grey lines. The residues shown as spheres are part of the 3 symmetric lactose binding sites while those shown as sticks are part of the single sodium binding site (Figure 5.2). The rotation indicated gives the view in E and F. (D) Contact map of ThreeFoil, with contacts deleted in MUT1 and MUT2 shown in red and blue, respectively. All deleted contacts are long-range (far from diagonal). (E) Long-range contacts (red sticks) of the β2-β3 loop residues (red spheres at Cα positions) deleted in MUT1. (F) For MUT2 the same number of contacts were deleted such that MUT1 and MUT2 have a very similar ACO. However, these contacts (blue sticks) are spread over the entire protein. (G) Folding free energies of ThreeFoil (black, same as in A), Symfoil (SymF; grey), a hybrid protein with the ThreeFoil contact map projected on the Symfoil backbone (HYB; gold), the two ThreeFoil mutants (MUT1; red, MUT2; blue) and Hisactophilin (His; green) are plotted in units of their respective folding temperatures ($k_B T_f$, Y axis) at their respective transition midpoints as a function of the fraction of their respective native contacts (X axis). The SymF, HYB, MUT1 and MUT2 free energy profiles have very similar barrier heights, in between those of the highly kinetically stable ThreeFoil and the much less stable His.

5.5.4 High chemical and protease resistances of ThreeFoil and other high ACO/LRO proteins

Extremely slow unfolding has been associated with the capacity to maintain native form and function under harsh conditions [246], such as high concentrations of protease [264, 271, 245] and detergent [245, 272]. Protease resistance of a classic extremely kinetically stable protein, α-lytic protease, has been proposed to originate from its large and highly cooperative unfolding energy barrier resulting in a rigid native conformation with limited local openings and consequently limited proteolytically susceptible regions [264]. Also, challenge by a high concentration of Sodium Dodecyl Sulfate (SDS) has been used extensively for direct evaluation of protein kinetic stability based on the ability of SDS to induce denaturation by trapping hydrophobic residues exposed during even transient unfolding events [272, 273]. Given its high barrier to unfolding, we tested ThreeFoil for resistance to protease and SDS. In the manner of Manning and Colón in their profiling of protein kinetic stability, we incubated ThreeFoil with the aggressive and non-specific protease, proteinase K [245]. ThreeFoil demonstrated strong resistance, remaining intact for the full 4 day challenge by a high concentration of protease (Figure 5.6A). A highly protease-resistant control protein, the dimeric human Cu,Zn superoxide dismutase (SOD) also remained intact, while hisactophilin which has greater thermodynamic stability but much faster unfolding kinetics than ThreeFoil (Table 5.1), was completely degraded within an hour as were other commonly studied proteins (Figure 5.6A). The results for the SDS challenge follow the same pattern with only ThreeFoil and SOD being resistant (Figure 5.6B), although SOD depends on an intact disulfide bond for SDS resistance while ThreeFoil does not (Figure 5.15F). Given the correlations between high topological complexity and slow unfolding (Figure 5.4A,B) [78] and between slow unfolding and SDS/protease resistance [246, 245], we asked if these resistances could be predicted from topological complexity. We conducted experiments and surveyed the literature to identify proteins with experimentally demonstrated resistance, or lack thereof, to SDS or protease. The identified proteins include new (Figure 5.16) and previously reported [271, 272] results from whole proteome screening to identify kinetically stable proteins, as well as new (Figure 5.6A,B, Figure 5.15) and previously reported analyses of individual proteins (Table 5.3). The results (Figure 5.6C,D) clearly show that resistant proteins have notably high ACO/LRO values, similar to ThreeFoil, whereas the non-resistant proteins tend to have much lower values. The few non-resistant proteins with a high ACO/LRO indicate that high topological complexity is necessary but not always sufficient for resistance. This suggests the rough measure provided by ACO/LRO does not capture other requirements such as highly cooperative unfolding, needed to eliminate weak points in the structure which provide opportunities for attack by chemical denaturants and proteases [264, 246, 245]. Thus, a high ACO/LRO indicates potentially high resistance to degradation/denaturation but a more detailed simulation, as performed for ThreeFoil (Figure 5.5), is needed for a more accurate prediction and understanding of molecular determinants for resistance. Finally, we note that the distribution of ACO/LRO values for a large dataset of proteins with previously characterized kinetics, similar to the non-resistant cases, is markedly lower than for the resistant proteins (Figure 5.6C,D). Thus, while kinetically stable resistant proteins exist, they have received relatively little attention and using their folds or incorporating analogous long-range contacts provides attractive new avenues for designing resistance.

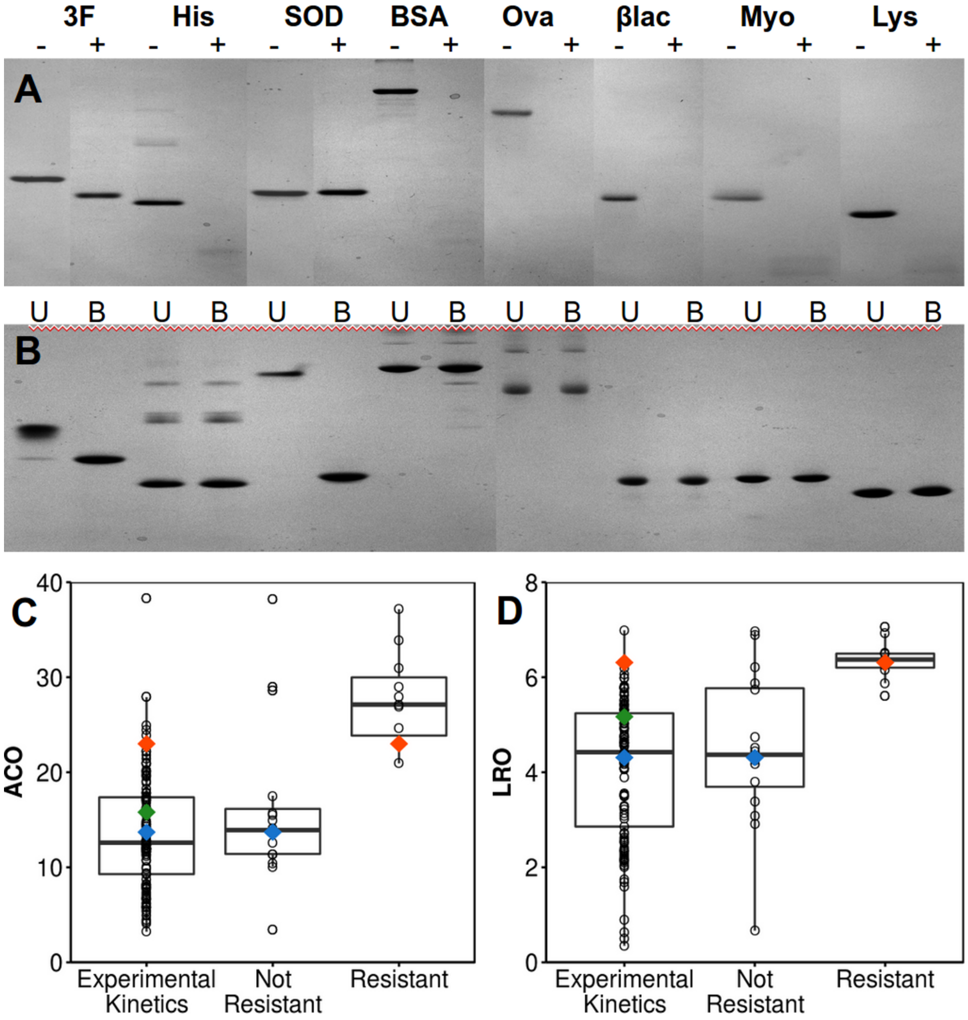


Figure 5.6: ThreeFoil is highly resistant to protease and detergent. (A) Proteins incubated with Proteinase K: ThreeFoil (3F), hisactophilin (His), human Cu,Zn superoxide dismutase (SOD), bovine serum albumin (BSA), ovalbumin (Ova), β-lactoglobulin (βlac), myoglobin (Myo), and lysozyme (Lys). Protein before (-) and after incubation with protease (+) are shown. Retention of the protein band after incubation shows resistance to digestion. ThreeFoil and SOD are shown after 4 days (still non-degraded), while others are shown after 1 hr (fully degraded). The molecular weight decrease for ThreeFoil after incubation is due to the loss of its unstructured His-tag (Figure 5.15A). Individual gels shown in Figure 5.15B-E. (B) The same proteins tested for resistance to SDS. Where the boiled (+) and unboiled (-) samples are indistinguishable, no SDS resistance is observed, while a higher running unboiled sample indicates SDS is unable to penetrate and bind without thermal unfolding of the protein. Comparison of topological complexity as measured by (C) ACO and (D) LRO for proteins that have been kinetically characterized experimentally (Table 5.2) and those with experimentally demonstrated resistance or non-resistance to protease and SDS (Table 5.3). Resistant proteins generally have higher topological complexity. β-trefoil proteins are colored as in Figure 5.4. Data shown as box-and-whisker plots, with a horizontal line at the median, box enclosing middle 50% of the data, whiskers drawn to 1.5*IQR (inter-quartile range).

5.6 Discussion

An in depth analysis of the folding characteristics of designed proteins, as we have performed for the 3-fold symmetric ThreeFoil, is rarely reported, yet is critical for ultimately understanding design outcomes and improving their reliability. We demonstrate a high level of design success for ThreeFoil as evidenced by its: 1) reversible, cooperative, two-state (un)folding, and 2) well folded and functional native structure which has high solubility and monodispersity, well-diffracting crystals, and great resistance against H/D exchange [44], denaturation by chaotropes and detergent, and degradation by protease. While the rational design of proteins with desired structure and function remains a great challenge and often require multiple cycles of design and/or selection to improve them, successes in designing both structures and/or functions, including ones not observed in nature, have been increasing [39, 254, 21, 14, 15, 47, 18, 274]. These results demonstrate the increasing understanding of fundamental principles and utility of computational protein design. Recently, there have been multiple reports of success for common folds based on repeated structural elements, including relatively high success rates and stabilities for various helix-containing elongated repeat proteins [47, 65] and toroidal or globular superfolds [44, 253, 39, 79, 74, 45, 50]. The great diversity of sequences observed for these symmetric protein structures may reflect an inherent capacity for stability, foldability and functionality that is especially amenable to both evolution and design [43]. Design strategies similar to that used to make ThreeFoil, which employ repetition of structural elements designed using existing sequence information and structural modelling with the Rosetta energy function [260], have proven particularly fruitful, with several studies yielding well-folded proteins with high melting points on the first attempt [44, 74, 45, 47]. Further, we have shown that ThreeFoil possesses stability, cooperativity and multivalent binding function. These features may be “inherited” through the use of existing sequence information, generating a more naturally funnelled energy landscape. Other proteins designed in a similar way, and not yet characterized in detail, may also capture favorable natural features [74, 45, 47]. Also, our results show how ligand binding can further smooth the landscape by decreasing the formation of non-native structure and so promote folding and design success. While evolution has provided a great range of sequences and structures that may be leveraged, it has also set limitations, which need not constrain rational protein design. As an example, natural proteins for which kinetics have been measured typically unfold on a timescale of seconds-hours [78]; ready unfolding may be needed to facilitate protein transport, regulation or turnover. However, other natural proteins that must withstand harsh extracellular or thermophilic conditions tend to have high kinetic stability [264, 246] hence fast unfolding is not an inherent constraint on proteins. Artificial proteins can be freed from various biological constraints allowing for uncommon properties such as extreme kinetic stability using suitable natural structures or novel ones with the requisite features. It is important to note that while the energy landscape of a protein defines both its thermodynamic and kinetic stabilities, the two properties are distinct. Thermodynamic stability depends on the energy difference between folded and unfolded states while kinetic stability depends on the energy barrier between the folded and transition states (Figure 5.3). High kinetic stability is a particularly attractive feature for rational design, as it is linked to other benefits such as resistances against protein denaturation and degradation by detergents (by decreasing exposure of the hydrophobic core), proteases (by limiting accessibility of cut sites), and temperature (by producing a high energy transition state barrier that is unlikely to be crossed by thermal fluctuations) [246, 245]. Such characteristics are highly desirable for industrial or biomedical applications that require a protein to remain folded and functional for a long time, even in challenging environments. While it is known that kinetic stability and its associated resistances are the result of slow global unfolding and limited local opening [264, 246, 271, 245], little has been reported on how to rationally incorporate this into designed proteins. Our in depth experimental and modelling analyses of ThreeFoil provide valuable insight into the molecular basis for these characteristics. Specifically, the origin of ThreeFoil's very slow global and limited local unfolding is a high and steep energy barrier which is a consequence of a folded topology that includes a large number and proportion of long-range and extensively distributed contacts. Thus, there are no weak points in the structure and it undergoes very cooperative folding to a native state that is highly resistant to local openings. In summary, a simple calculation of ACO/LRO indicates whether a design has the capacity to be kinetically stable, while Gō model simulations give a more accurate prediction and can be used to determine the impact of specific contacts. This paves the way for rational design of resistance to harsh conditions. The mechanistic understanding of the structural determinants of resistance and the ability to design it, as well as the simplified and efficient design process of using structural repetition within the context of a symmetric and functional superfold, provide valuable avenues for improving future protein designs.

5.7 Methods

5.7.1 Expression and purification of ThreeFoil

ThreeFoil was expressed from a pET-28a plasmid in BL21 DE3 *E. coli* cells. Expression was induced (1 mM IPTG) and cells were harvested after 48 hours at 37 °C. Inclusion bodies were solubilized in urea (6 M urea, 100 mM sodium phosphate, 10 mM Tris at pH 8.1), bound to a Ni-NTA column, and eluted at pH 4.5. The protein was then refolded by dialysis in standard buffer (100 mM sodium phosphate and 300 mM NaCl at pH 6.6). To remove bound sodium, purified protein solution was dialyzed 1:10 against milliQ-H_2O for 4 hours, repeated 10 times, concentrated by stirred cell (Amicon) ultrafiltration (10 kDa, Millipore) and lyophilized.

5.7.2 Kinetic measurements

All measurements performed at 27 °C. Manual mixing induced unfolding of ThreeFoil was performed by addition of denaturant (GuSCN, GuHCl, urea) in standard buffer, and refolding was induced by addition of standard buffer to unfolded ThreeFoil (either 4 M GuSCN, 6 M GuHCl, or 6 M urea, in standard buffer). All final protein concentrations for kinetic measurements were 3.3 μM. Unfolding and refolding kinetics at different denaturant concentrations were monitored by fluorescence or circular dichroism (CD) (see SI Materials and Methods). Both continuous and discontinuous fluorescence measurements in GuHCl and urea (Figure 5.7C) were obtained using a 1 cm pathlength quartz cuvette, with excitation at 280 nm and a slit-width of 1 nm, and emission monitored at 313 nm with a slit width of 3 nm, using a Spex® Flourolog-311 (Horiba) fluorimeter. Samples (1 mL) were stored in Eppendorf tubes sealed with parafilm in an incubator and measured periodically in the discontinuous case. Continuous fluorescence measurements using GuSCN (Figures 5.3, 5.7) were monitored by fluorescence, with excitation at 274 nm and emission at 317 nm using a SpectraMax M5 plate reader (Molecular Devices). To folded or unfolded (standard buffer or 4 M GuSCN in standard buffer, respectively) protein (550 μM, ∼10 mg/mL), standard buffer with varying concentrations of GuSCN was added (3.3 μM final protein concentration). Fluorescence was measured from the bottom of the plate using 96-well UV Star® (Greiner Bio-One) black-well plates with clear UV transparent bottoms and tops covered with HD Clear sealing tape (Hampton Research) to prevent evaporation. The total run time (30 minutes to 4 days) and interval of readings (10 seconds to 30 minutes) were chosen such that each run would have ∼200 measurements. Each trace was fit to a single exponential equation:

$$S = B + Ce^{-kt} \tag{5.1}$$

where S is the fluorescence signal, B is the offset, C is the amplitude, k is the rate constant, and t the time in seconds. The chevron for the observed rate constant, k_{obs}, as a function of denaturant activity, A, was fit to an equation for a two-state transition between the folded ($_f$) and unfolded ($_u$) states of the protein [221]:

$$\ln\left(k_{obs}\right) = \ln\left(e^{F+m_f A} + e^{U+m_u A}\right) \tag{5.2}$$

where m_f and m_u are the linear denaturant-dependence of folding and unfolding, respectively, and F = ln(k_f^{H2O}) and U = ln(k_u^{H2O}) are the natural logarithms of the respective folding and unfolding rate constants in water (in this case measured in s^{-1}). Activity (A) was calculated using [275]:

$$A = [GuSCN]\left(\frac{C_{0.5}}{C_{0.5} + [GuSCN]}\right) \tag{5.3}$$

where $C_{0.5}$ is the [GuSCN] at half activity, 6.47 M [275]. Kinetic data was fit to Eq. 5.2 using the Origin software (OriginLab), and uncertainty values obtained from the fit are reported in Table 5.1. The m-value is calculated as:

$$m = m_u - m_f \tag{5.4}$$

and reflects the total increase in solvent accessible surface area in going from the folded to the unfolded state. The β-Tanford value (β_T) is a measure of the change in solvent-accessible surface area of the transition state, and is calculated as:

$$\beta_T = -m_f/m \tag{5.5}$$

with a value of 1 indicating a native-like transition state and a value of 0 indicating an unfolded-like transition state. The Gibbs free energy of unfolding or the thermodynamic stability of the protein, is calculated from the equilibrium ratio of unfolded to folded protein concentrations given by:

$$\Delta G_{U/F} = -RT\ln\left(k_u^{H20}/k_f^{H2O}\right) \tag{5.6}$$

where R is the universal gas constant (0.001987 kcal mol^{-1} K^{-1}), and T, the temperature (300.1 ± 0.5 K for ThreeFoil, 298.1 K for Symfoil). A larger value indicates a higher stability against unfolding.

5.7.3 SDS resistance

Protein in H_2O was diluted into sodium dodecyl sulfate (SDS) and Tris so that final samples contained 0.5 mg/mL protein and 1% SDS in 125 mM Tris (pH 6.8). Samples were then either boiled or incubated at room temperature for 10 min prior to analysis by SDS-PAGE using 15% Acrylamide gels with 0.1% SDS in Tris/glycine running buffer (pH 8.3), and either without (Figure 5.6B) or with (Figure 5.15F,G) 7% (v/v) β-mercaptoethanol to reduce disulfides.

5.7.4 Protease resistance

Samples contained 0.5 mg/mL of protein in 25 mM Tris and 1 μM EDTA (pH 8.3). A time zero control was taken before adding proteinase K (final concentration 0.02 mg/mL), and further samples taken after 1 hr, 1 day, and 4 days of incubation at 25 °C. The reaction was stopped by mixing samples 1:1 with buffer (2.5 μM phenylmethylsulfonyl fluoride, 125 mM Tris, 4% SDS (w/v), 20% (v/v) glycerol, 15% (v/v) β-mercaptoethanol, at pH 6.8) and boiling for 10 min. SDS-PAGE was performed using the same gel conditions as for SDS Resistance.

5.7.5 Carbohydrate binding affinity

Sugar binding affinity was measured by fluorescence using a plate-reader with settings as for the kinetic measurements. To each well, 200 μL of protein solution (4.16 μM protein in standard buffer was added, followed by 50 μL of carbohydrate solution (varying concentrations of carbohydrate in standard buffer), yielding a final protein concentration of 3.3 μM. Samples were then equilibrated for 30 minutes at 27 °C and then measured in quadruplicate. No change in fluorescence signal was observed upon addition of sucrose, and so the data were not fit to a dissociation constant. For lactose the dissociation constant, $K_{d,F}$, was determined by fitting; the fluorescence data to a hyperbolic equation for binding to 3 identical, non-interacting sites:

$$S = \frac{S_{max}[L]}{[L] + K_{d,F}} \tag{5.7}$$

where S is the measured change in fluorescence, S_{max} is the difference in fluorescence between the protein with 3 bound ligands and the free protein, and $[L]$ the concentration of ligand. The fit yielded a $K_{d,F}$ of 1.22 ± 0.15 mM. The apparent change in stability of the protein native state resulting from ligand binding ($\Delta\Delta G_{binding}$) was calculated as:

$$\Delta\Delta G_{binding} = -RT * \ln(K_{U-F,app}) \tag{5.8}$$

where $K_{U-F,app}$ is the equilibrium constant between denatured and native protein and is calculated as:

$$K_{U-F,app} = \frac{\sum[UL_{0-3}]}{\sum[FL_{0-3}]} = \frac{[U]}{[F]} * \left(\frac{1 + \left(\frac{[L]}{K_{d,U}}\right) + \left(\frac{[L]}{K_{d,U}}\right)^2 + \left(\frac{[L]}{K_{d,U}}\right)^3}{1 + \left(\frac{[L]}{K_{d,F}}\right) + \left(\frac{[L]}{K_{d,F}}\right)^2 + \left(\frac{[L]}{K_{d,F}}\right)^3} \right) \tag{5.9}$$

where $[UL_{0-3}]$ and $[FL_{0-3}]$ are the concentrations of unfolded and folded protein, respectively, with any number of ligands bound (from 0 to 3). $K_{d,U}$ and $K_{d,F}$ are the dissociation constants for the identical, independent, non-interacting binding site in the unfolded and folded protein, respectively. Using the known $K_{d,F}$ of 1.22 ± 0.15 mM, a range for $\Delta\Delta G_{binding}$ was calculated for $K_{d,U}$ in the range of 50 to 1000 mM. Given these ranges for $K_{d,U}$ and the uncertainty in $K_{d,F}$, the expected stabilization from lactose binding ($\Delta\Delta G_{binding}$) ranges from 5.6 to 6.9 kcal/mol, which agrees well with the observed stabilization from kinetic measurements (Table 5.1) of 7.0 ± 0.2 kcal mol^{-1}. Glycan array analysis (Figure 5.9) was performed as reported by the Consortium for Functional Glycomics [177] (www.functionalglycomics.org/static/consortium/resources/resourcecoreh.shtml), using a ThreeFoil concentration of 0.2 mg/mL.

5.7.6 Expected values of m in GuSCN

The expected m-value of ThreeFoil in GuSCN was calculated from the expected value in urea [276] multiplied by a factor of 7.46, which is the average scaling-factor from comparison of several protein m values between urea and GuSCN [275]. The reported m, mf and mu values for Symfoil in GuHCl [79] were converted to estimated values in GuSCN by first dividing by a factor of 2.37 to obtain the expected value in urea [276], and then multiplying by 7.46 as above.

5.7.7 Circular dichroism measurements

CD was measured using a J715 spectropolarimeter (Jasco) at room temperature. For kinetic measurements a 0.1 cm pathlength quartz cuvette was used with measurements at

230 nm. ThreeFoil concentration was 3.3 uM in standard buffer with either 0.84 M urea or 0.05 M GuHCl. For measurements of spectra, a 0.02 cm pathlength quartz cuvette, 0.1 nm step size, and 50 nm/min scan rate were used. Protein concentration was 5.5 μM in potassium fluoride buffer (300 mM KF and 100 mM sodium phosphate, pH 6.6) with or without 6 M urea. Potassium fluoride was used instead of sodium chloride to decrease the absorbance due to chloride ions and so allow for secondary structural analysis. High absorbance values at high denaturant concentrations preclude quantitative secondary structure analysis of denatured samples. However, quantitative analysis was possible for ThreeFoil and OneFoil in the absence of denaturant, and was performed using Dichroweb [277, 278] and the CDSSTR algorithm [279] with dataset-7 [278].

5.7.8 ANS binding

Solutions were made to a final protein concentration of 1 μM and an ANS concentration of 150 μM for an ANS:amino acid ratio of ∼1:1. Stock ANS was filtered (0.45 μm syringe filter, Pall acrodisc) and concentration (10 mM) determined by absorbance at 350 nm using a molar extinction coefficient [280] of 5000 $M^{-1}cm^{-1}$. All samples were in standard buffer, with denatured samples also including 6 M urea, and carbohydrate-containing samples including lactose or sucrose at 50 mM. Fluorescence was measured using a 1 cm pathlength quartz cuvette with excitation at 360 nm and a slit-width of 1 nm, and emission monitored from 420-600 nm at 1 nm intervals with a 3 nm slit-width using a Spex® Flourolog-311 (Horiba) fluorimeter at room temperature.

5.7.9 NMR

2D ^{1}H-^{1}H NOESY spectra were obtained for ThreeFoil samples at 27 °C using a Bruker Avance DMX 600 MHz spectrometer with a TSI probe and excitation sculpting for water suppression [176]. The NOESY mixing time was 125 ms. The initial protein concentration was 1.1 mM (20 mg/mL) in standard buffer containing 7% (v/v) D_2O, to which concentrated lactose solution (450 mM in native buffer) was added. Spectra were obtained for 0, 0.32, 1.28, 4.80, and 12.8 mM lactose concentrations and were processed using XWinNMR (Bruker).

5.7.10 Sample preparation and diagonal 2D (D2D) SDS-PAGE

Processed lysate was concentrated to 150µL and incubated for 5 min in SDS sample buffer (pH 6.8) to a final concentration of 45 mM Tris HCl, 1% (w/v) SDS, 10% (v/v) glycerol, 0.01% (w/v) bromophenol blue at room temperature. Sample was loaded without prior heating into a well of a 12% acrylamide gel (16 cm x 14 cm x 1.5 mm). Electrophoresis was performed in a Protean II xi cell (Bio-Rad, Hercules, CA) by using 480 V and 50 mA. The gel was kept at 10 °C by using a circulating water bath. Running buffer contained 25 mM Trisaminomethane (Tris base), 0.2 M glycine, and 0.1% (w/v) SDS. After the first-dimension run, the gel strip was cut out and incubated for 10 min in equilibration buffer (45 mM Tris HCl, 1% (w/v) SDS, 10% (v/v) glycerol, 0.01% (w/v) bromophenol blue, pH 6.8) at 98°C. The gel strip was drained briefly and placed on top of a 12 cm x 14 cm x 2 mm 12% acrylamide gel. A small amount of 12% acrylamide solution was used to re-polymerize and fuse the strip to the resolving gel. The second-dimension separation was performed under similar conditions as the first-dimension run except 65 mA was used for each gel. Gels were stained with Coomassie blue (Bio-Rad Biosafe). Destained gels were imaged by a Biorad Gel Doc XR+ system.

5.7.11 Protein identification and mass spectrometry

Protein spots below the gel diagonal protein bands were picked by using a One Touch 2D gel spotpicker (1.5 mm), and then digested with trypsin. (Promega, Madison, WI).

For MALDI/TOF, α-Cyano-4-hydroxycinnamic acid (CHCA) solution was used as matrix. External mass calibration was performed by using peptide standards with a mass range of 700–3200 Da (Bruker Dalton, Germany). The protein was identified by matrix assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF/MS) (Autoflex III Brucker Daltonics, Germany). The parameters for MALDI-TOF/MS analysis were set as follows: reflection mode, trypsin as the digestion enzyme with one cleavage site, variable modifications oxidation (M), carbamidomethyl (C), peptide mass tolerance ± 100 ppm. Results were analyzed using BioTools (Bruker Daltonics, Germany), and then were searched against NCBInr database without specify taxonomy using MASCOT search engine (Matrix Science Ltd.). Only those proteins identified by MASCOT search criteria with significant scores ($P<0.05$) and taxonomy *Thermus thermophilus* were considered acceptable.

For LC/MS/MS, the resulting peptide mixture was analyzed using an Agilent 1200-Series LC system coupled to an LTQ-Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The LC system was equipped with a 75 µm ID, 15 µm tip, 105 mm picochip (New Objective, Cambridge, MA) bed packed with 5 μm BioBasic, (Thermo Scientific, Bremen, Germany) C18, 300A resin. Sample loading was finished in 2% buffer B (98% ACN in 0.1% formic acid) in 10 min. Elution was achieved with a gradient of 15-90% buffer B in 75 min. The flow rate was passively split from 0.3 mL/min to 200 nL/min. The mass spectrometer was operated in data-dependent mode to switch between MS and MS/MS. The six most intense ions were selected for fragmentation in the linear ion trap using collisionally induced dissociation. Mass spectrometry data obtained from all LC-MS-MS analysis were searched against Swissprot using Sequest search algorithms through Proteome Discoverer (Thermo Scientific, Bremen, Germany). Enzyme specificity was set as trypsin with maximum three missed cleavage allowed. Carbamidomethylation of cysteine and oxidation of methionine were included as variable modifications. The mass error of parent ions was set to 10 ppm and 0.8 Da for fragment ions. Commonly accepted criteria for high-confidence peptide identifications (xCorr 1.8 for +1, 2.5 for +2, 3.5 for +3) was used to screen peptides. To avoid false positive protein identification, each protein included in the results table contains more than 2 high-confidence unique peptides.

5.7.12 *Ab initio* folding of OneFoil using Rosetta

The sequence of OneFoil:
GDGYYKLVARHSGKALDVENASTSDGANVIQYSYSGGDNQQWRLVDL,
was used with the Rosetta *ab initio* folding program [260] to generate 100,000 predicted structures. The lowest energy structure was a simple β-sheet comprised of 4 anti-parallel strands. This was also the most common topology, accounting for over a third of all predicted structures as judged by a TM-score [281] of greater than 0.5. These simluations were made possible the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada.

5.7.13 Details of the coarse-grained Gō models

The model. We use a coarse-grained model of the protein which represents each amino-acid by its Cα atom. The potential energy function in this protein model includes only those interactions which are present in the native structure of the protein (the X-ray or NMR structure). Usually, no non-native interactions are included. Such structure-based or Gō models encode the funneled energy landscapes of proteins and have been used to understand protein folding because they correctly capture folding mechanisms and trends in rates while being computationally inexpensive. Here, we use a previously extensively tested (and used) form of the Gō model, whose details have been previously reported [282, 236], to perform molecular dynamics (MD) simulations.

Contact maps of wild-type (WT) ThreeFoil, Symfoil and Hisactophilin. The two inputs required to define the energy function of the Gō model are the coordinates of all Cα atoms, which can be extracted from the PDB file, and the contact map. The contact map defines all possible attractive interactions that are present between the Cα beads. Residues i,j are only considered to be in contact if |i-j|>3. Further, all contacts are energetically equivalent unless specified otherwise. Here, we use the contacts of structural units (CSU) analysis [283] to find contacts between the atoms in the folded structures of the proteins. These atomic contacts are then projected back onto their respective Cα beads. The interaction energy of a contact is at its minimum when the distance between the two Cα atoms which participate in the contact equals the distance between these atoms in the folded state. We find that ThreeFoil (PDB: 3PG0) has 140 Cα atoms with 410 contacts, Symfoil (PDB: 3O4D) 123 Cα atoms with 359 contacts and Hisactophilin (PDB: 1HCD) 118 Cα atoms with 324 contacts. The ACO values of the proteins calculated from their Cα contact maps are: ThreeFoil: 38.54, Symfoil: 32.64 and Hisactophilin: 28.66.

Contact maps of HYB. We also simulate a hybrid (HYB) which has the ThreeFoil contact map projected onto the Symfoil backbone. To create the energy function of HYB we use the coordinates of the Symfoil. To calculate the contact map we first structurally align the two proteins using the Multiseq extension [284] of VMD [285]. We then use this alignment to pick those contacts of ThreeFoil both of whose participating residues (Cα beads) have corresponding residues in Symfoil. The HYB has 123 Cα atoms with 352 contacts. The ACO of HYB is 32.51.

Contact Maps of MUT1 and MUT2. The coordinates of the Cα atoms for both of these mutants are extracted from the coordinate file (3PG0) and are the same as for WT ThreeFoil. Therefore, the bond, angle and dihedral terms of the energy function are the same as WT and consequently local structural propensities are the same. All long ranged contacts (contacts between residues i and j where |i-j|>11; shown in red in Figure 5.5D and as red sticks in Figure 5.5E) of the residues (19-26, 67-73, 114-120) (shown as red spheres in Figure 5.5E) are deleted to create MUT1. The following contacts, shown in Figure 5.5D,F, were deleted to create MUT2: ((2,57), (4,41), (4,55), (4,57), (5,40), (5,41), (5,42), (7,110), (9,112), (10,96), (10,98), (11,112), (14,32), (16,54), (16,123), (18,56), (18,78), (27,77), (27,78), (28,77), (29,78), (29,122), (30,121), (30,122), (31,112), (31,121), (31,123), (31,136), (34,141), (41,141), (42,78), (47,104), (48,104), (49,104), (51,88), (51,102), (51,104), (52,88), (61,79), (65,125), (74,124), (74,125), (75,124), (76,125), (89,125), (99,135), (100,136), (101,136), (101,138), (106,125), (108,126), (108,127), (109,128)). These contacts do not overlap with the contacts deleted to create MUT1 but were randomly chosen from other long ranged contacts such that both proteins have a similar ACO: MUT1: 35.82 and MUT2: 35.83. MUT1 and MUT2 have a higher ACO than HYB because of the longer protein length. They have 140 Cα atoms with 357 contacts.

Symmetrized contact map. Usually, no non-native interactions are included in Gō models. However, ThreeFoil is formed by a 3-fold repeat of a 47 amino-acid sequence (termed OneFoil). Consider a contact i_x and j_y which is present in the native state of the protein. Here, 1≤i,j≤47 is the position of the residues within a repeat while 1≤x,y≤3 is the index of the repeat that they are present in. An energetically equivalent contact may be formed between iX and jY while folding, where X and Y denote repeats other than x and y. In order to understand the effects of such contacts, we simulate a symmetrized model of ThreeFoil where if a contact (i_x,j_y) is present in WT then all contacts (i_X,j_Y), where X and Y assume all values between 1 and 3, are present in the contact map. The minimum interaction energy distance of a symmetrized contact is set to be the same as the minimum interaction energy distance of the corresponding WT contact. The weights of the symmetrized contacts are the same as the weight of the WT contacts). Such symmetrized Gō models have previously been used to study domain swapping and repeat proteins [286, 287, 263]. The ThreeFoil symmetrized contact map has 1217 contacts and is shown in Figure 5.13A.

Non-native contacts from the most populated Rosetta structure of OneFoil. In order to understand the effect of possible residual non-native structure on the folding of ThreeFoil, we calculate the contact map and interaction distances from the structure of OneFoil shown in Figure 5.14A. This structure, henceforth called Rosetta OneFoil, is a representative structure chosen from the most populated cluster obtained from Rosetta [260] *ab initio* folding simulations of the OneFoil. From the calculated contact map we first choose those contacts which are not present in the contact map of the WT OneFoil (Fig S8B). We further filter these contacts by choosing only those whose contact distance in the WT OneFoil structure is 2.1 times the distance in the Rosetta OneFoil. This distance is such that about half of all the contacts from Rosetta OneFoil are chosen. These contacts, henceforth referred to as Rosetta contacts, are added to the WT OneFoil contact map with a minimum interaction energy distance calculated from the Rosetta OneFoil structure (Figure 5.14A; grey contacts). In order to choose a weight for the Rosetta contacts, we performed simulations of the OneFoil using a Gō model created from the WT Cα coordinates and the combined WT and Rosetta contacts. Upon varying the weights of the Rosetta contacts, we found that simulations which use a Rosetta contact weight of 1.5 times the WT contact weight populate both the Rosetta OneFoil and the WT structures almost equally (Figure 5.14B). Based on these simulations we choose this Rosetta contact weight. This recipe for creating a Gō model with two input structures has been used earlier to study conformational transitions [288]. For ThreeFoil simulations, the weighted Rosetta contacts are added to the WT ThreeFoil contact map within each of the three repeats. This contact map has 512 contacts and is shown in Figure 5.14A.

Molecular dynamics (MD). Folding simulations were performed at the folding temperature (T_f). T_f is the temperature at which the equilibrium population of the folded and the unfolded ensembles is equal. At T_f several transitions are obtained between the folded and the unfolded ensembles and this ensures adequate sampling of the transition region. This condition is similar to the transition midpoint in folding experiments where denaturant rather than temperature is used to achieve equally populated folded and unfolded ensembles. Further details about comparing results from T_f to those from transition midpoint are given in the Supplemental Methods. The folding temperatures of the studied proteins in the same energy units as the native contact weights are: ThreeFoil, 1.12; SymFoil, 1.13; HYB, 1.12; MUT1, 1.04; MUT2, 1.01; Hisactophilin, 1.12. Folding temperatures for proteins presented in the Supplemental Methods: ThreeFoil with symmetrized non-native interactions: 1.19; ThreeFoil with Rosetta non-native interactions: 1.12.

Because kinetically stable proteins are by definition slow to fold and unfold the waiting time between transitions gets longer and it becomes computationally expensive to simulate even a few transitions at T_f. In order to achieve extensive sampling in a reasonable amount of computer time we use an enhanced sampling technique called the modified multicanonical method [289]. In this method, the population of the transition region between the folded and the unfolded states is promoted by rescaling the MD force by an appropriately tuned Gaussian weight. These trajectories were then reweighted to obtain the usual canonical distribution. This technique was used to enhance the sampling in all simulations except those of OneFoil and reweighting was performed prior to all analyses of such simulations.

Analysis of all simulations. Since one of the inputs of the Gō model is the contact map, any function of the total number of native contacts becomes a natural coordinate that measures how folded a protein is. Here, we use a common reaction coordinate, the fraction of formed native contacts, Q [226, 282, 236]. Q varies from 0 to 1 with the protein being more folded at larger Q and less folded at smaller Q. We investigate the nature and mechanism of folding by calculating various quantities such as contact maps, the free energy ($\Delta\mathrm{G}/k_BT_f$), the radius of gyration (R_g), etc., as functions of Q. The different proteins that we simulate have different numbers of contacts and different folding temperatures. Using Q (fraction instead of number of contacts formed) and scaled folding free energies allows us to compare these proteins.

To calculate the transition state structure shown in Figure 5.5B,C, all protein snapshots which had $0.3<Q<0.4$ are chosen. We then determine the number of snapshots in which a given contact (between say residues i and j) is formed and divide this by the total number of snapshots. This gives the average formation or Q_{ij} ($0<Q_{ij}<1$) of individual contacts which in turn gives the average contact map [226, 282, 236]. The “foldedness” or formation of a given residue is calculated from this average contact map by averaging over the formation of all the contacts that the residue is a part of.

Analysis of simulations with added non-native contacts. We first calculate the free energy profiles of the proteins with non-native contacts as a function of native Q and compare them to that of ThreeFoil (no non-native contacts; Figure 5.5A). The free energy barrier height of the protein model with symmetrized contacts is smaller than that of ThreeFoil (Figure 5.13B). This is a known effect [241] caused by a shrinking in the size of the unfolded states of the protein upon addition of any attractive interaction between Cα beads. This shrinking promotes native contact formation and reduces the barrier. In order to provide further evidence that this shrinking is likely to be the cause of the lowered barrier, we plot the radius of gyration (R_g) as a function of Q and confirm that the unfolded ensemble and transition state ensembles indeed have a smaller size (Figure 5.13B; dashed lines). The free energy profiles of the two non-native simulations (with symmetrized: Figure 5.13B and Rosetta contacts: Figure 5.14C; grey curves) show no significant dips in the free energy profiles as could be expected from the formation of any additional intermediates with non-native interactions.

In order to further analyze non-native structure formation, we calculate the number of non-native contacts formed as outlined below. Of the many non-native contacts included in the model only some are formed in any given snapshot of the protein and we calculate this number. As before, we bin the snapshots by the fraction of native contacts, Q and observe the number of non-native contacts formed on an average at any given Q. This is plotted in black in Figure 5.13C and Figure 5.14D. Once averaged over all snapshots at a given Q, each non-native contact is formed at a fractional strength $0<Q_{nn}<1$ and it is important to understand how strongly the contacts are formed. For instance, if specific non-native structure is stabilized then contacts present in that structure may be formed at greater strength than the average formation of the native contacts, Q, or the “foldedness” of the protein. The number of such contacts (with $Q_{nn}>Q$) are plotted in Figure 5.13D and Figure 5.14D in red as a function of how folded the protein is. This plot drops close to zero for both protein models even when the proteins are only 15% folded (Q$\sim$0.15). Comparing the black and the red lines we conclude that most non-native contacts are present at strengths lower than Q. This conclusion is further strengthened when we plot the number of non-native contacts completely formed (Q_{nn}=1) on an average at a given Q (green line). This average is calculated using:

$$\sum_{i} Q^{i}_{nn}\left(Q\right) \tag{5.10}$$

where $0 \leq Q^{i}_{nn}\left(Q\right) < 1$ is the average strength of the i$^{\text{th}}$ non-native interaction at a given Q; $0 < i < N_{nn}$ where N_{nn} is the total number of non-native contacts. The number of non-native contacts completely formed on an average is less than 5% of N_{nn} for both proteins (green line; Figure 5.13C and Figure 5.14D) while the total number of non-native contacts formed at any strength (black line; same figures) is much higher. Thus, at higher Q most non-native contacts although formed are formed at strengths much lower than Q. Finally, we compare the non-native contact formation in the proteins with stabilized non-native contacts (symmetrized; Figure 5.13D and Rosetta; Figure 5.14D; blue lines) with the non-native contact formation in WT ThreeFoil (no non-native contacts stabilized). The amount of non-native contact formation is first calculated for both the symmetric and the Rosetta contacts using the native-only Threefoil simulations. Then the root mean square deviations ($RMSD$) between the non-native contact lists of the two non-native models and the native-only model are calculated. The bare $RMSD$ is multiplied by the total number of non-native contacts and this gives the average number of extra non-native contacts that form in the proteins with the stabilized non-native interactions. The final $RMSD$ is calculated as:

$$RMSD = N_{nn} * \sqrt{\sum_{i} \frac{\left[Q^{i}_{nn}\left(Q\right) - Q^{i}_{nn}\left(Q\right)\right]^{2}}{N_{nn}}} \tag{5.11}$$

where $0 \leq Q^{i}_{nn}(Q), Q^{i}_{nn,Nat}(Q) < 1$ is the average strength of the i$^{\text{th}}$ non-native interaction at a given Q in the non-native model and the native-only model, respectively, and $0 < i < N_{nn}$ where N_{nn} is the total number of non-native contacts.

When the green and the blue lines are similar, non-native contacts are mostly formed only in the protein with the stabilized non-native interactions. However, some non-native contacts in the symmetrized non-native model are compatible with the folded state of the protein and can also be formed in the native-only model simulations. When many such contacts are formed the $RMSD$ reduces and the values of the blue line are smaller than the values of the green line (Figure 5.13D). In such instances, the blue line is a better representative of the average number of non-native contacts formed than the green line. The Rosetta non-native contacts are by definition not compatible with the native structure of the protein and in Figure 5.14D, the green and the blue lines are very similar.

Finally, the native (Figure 5.13E and Figure 5.14E) and non-native contact maps (Figure 5.13F and Figure 5.14F) of the two proteins with non-native contacts are plotted close to the transition state (Q ∼0.35; dotted lines in Figure 5.13D and Figure 5.14D), in order to observe the location of the highly formed non-native contacts. For the protein with the symmetrized non-native contacts, the highly formed non-native contacts are close (in the protein structure) to the native contacts as is expected from the comparison of the blue and the green curves in Figure 5.13D. The non-native interactions are very weakly formed in the protein with Rosetta contacts (Figure 5.14F). This is because the non-native interactions are not compatible with the native structure and can only form when there is little or no structure. Further, these intra-trefoil interactions are local and are only formed when there is little native structure within a given trefoil. The slight changes in folding routes seen in Figure 5.5B, Figure 5.13E and Figure 5.13F can be explained by the fact that the β-trefoil fold can accommodate several folding routes and minor changes in contacts has already been shown to cause a switch in folding routes [236, 289].

Using the free energy profiles (Figure 5.13B and Figure 5.14C), the various quantities of non-native contacts explained earlier (Figures 5.13C, 5.13D and 5.14D) and the native and non-native contact maps at Q∼0.35 (Figures 5.13, 5.13F, 5.14E, 5.14F), we conclude that there is little specific non-native contact formation and so non-native trapping is unlikely to be the primary cause of slow-folding in Threefoil.

Comparing folding barrier heights obtained at T_f with barrier heights obtained at transition midpoints. It has been shown that the barrier heights calculated from the specific Gō model used in our simulations can be directly correlated with temperature denaturation (or melting) experiments of proteins [226]. Here, we examine the correlation between folding barrier heights obtained in denaturant-induced equilibrium folding/unfolding experiments with those obtained from temperature-induced folding/unfolding experiments. We plot log(k_f) at denaturation midpoint, C_{mid}, vs. log(k_f) at the folding temperature, T_f, where k_f is the folding rate and k_f=k_u in both these conditions (Figure 5.12). We find that the rates are linearly correlated with a slope of 1.14. Thus, if one were to calculate the difference between the barrier heights of two proteins at T_f (as found by simulations) then that number needs to be multiplied by 1.14 to get the difference in barrier heights of the two proteins at C_{mid} (as typically found experimentally).

We find from simulations (Figure 5.5G) that the difference in the barrier heights (at the top of the barrier) for ThreeFoil and Symfoil (in units of k_BT_f as in work by Chavez and co-workers [226]) is 5.6 (17.9-12.3), when computed as the natural logarithm. By using the T_f to C_{mid} conversion factor, we expect the differences in experimental barrier heights at denaturant induced midpoints to be about 6.4 (5.6*1.14). We find that the difference in experimental barrier heights between the two proteins is 6.3 (17.1-10.8). Thus, the barrier height difference obtained from simulations (6.4) is remarkably close to that obtained from experiment (6.3). This comparison should be taken with a pinch of salt as both experimental and simulation barriers are accompanied by error bars which have not been taken into account here. Nevertheless, the calculation highlights that almost all of ThreeFoil's slow folding can be accounted for by its topology.

5.8 Supplemental Information

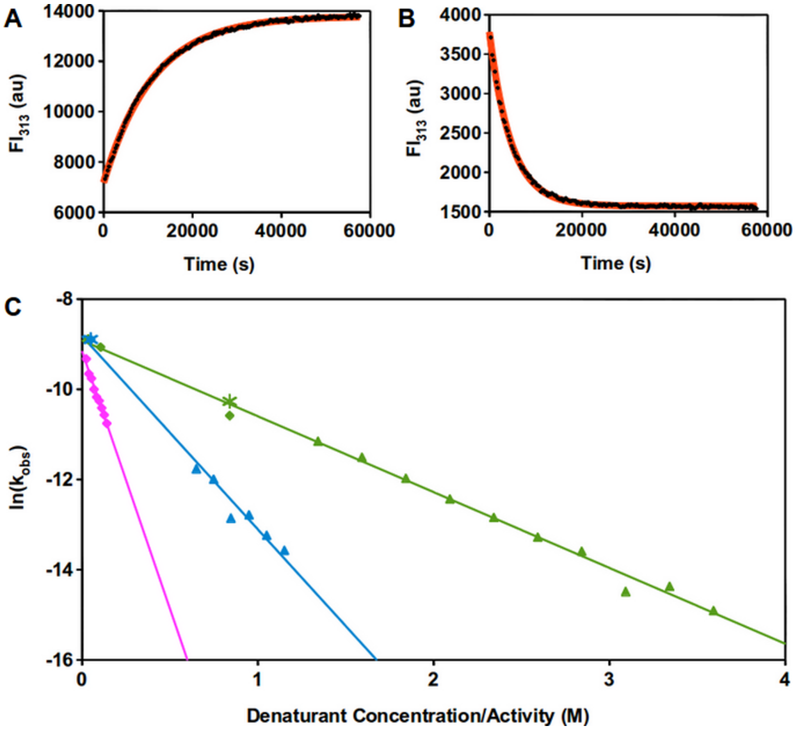


Figure 5.7: Reversible unfolding of ThreeFoil fits a two-state transition. Representative kinetic traces for ThreeFoil folding (A) and unfolding (B) in 0.02 M and 3.49 M GuSCN (black diamonds), respectively, show an excellent fit to a single exponential function (orange line, see Methods). (C) Observed rate constants (k_{obs} in s^{-1}) for folding measured in various denaturants (GuSCN - magenta, GuHCl - blue, and urea - green), extrapolate to the same k_f H_2O value in the absence of denaturant (solid lines). Rate constants measured using different optical probes (CD - stars, fluorescence - diamonds and triangles) are also in agreement. In addition, both discontinuous kinetic measurements performed over months (triangles) and continuous kinetic measurements performed over days (diamonds) give consistent rate constants. The single exponential fits and agreements of rate constants are characteristic of a two-state (un)folding transition.

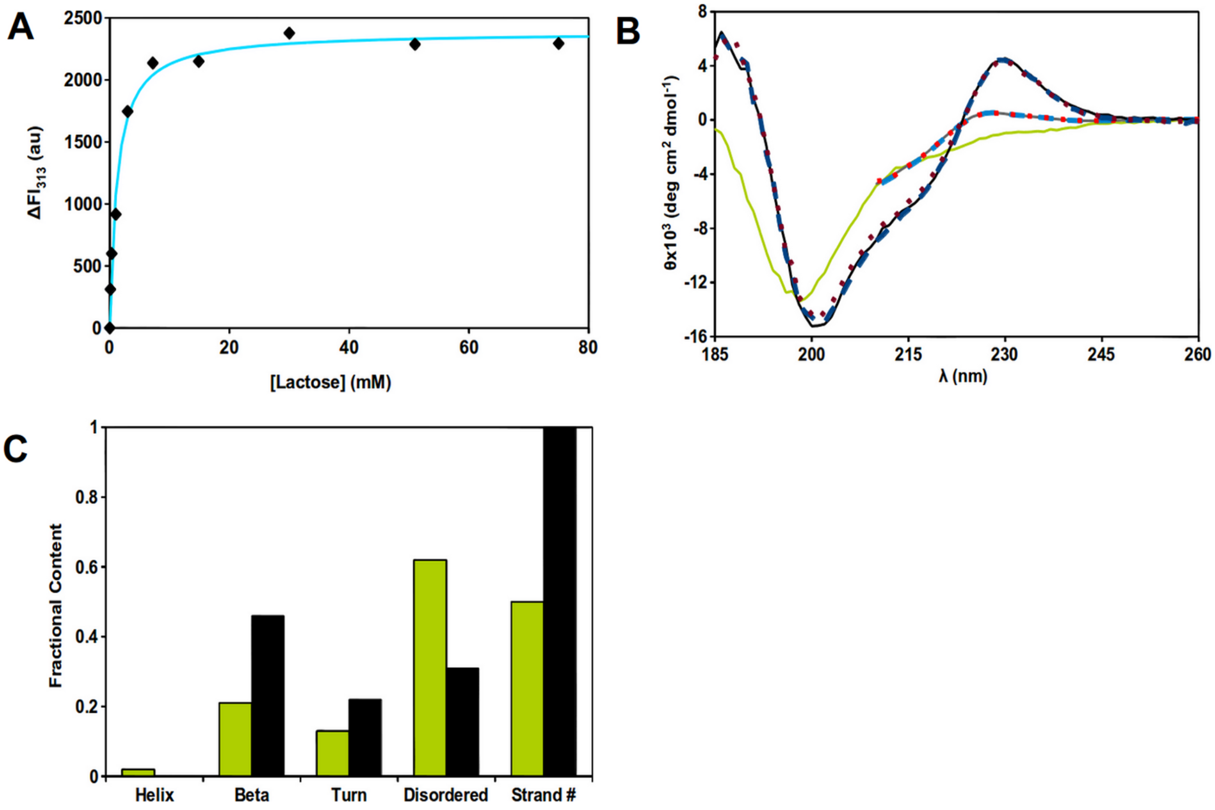
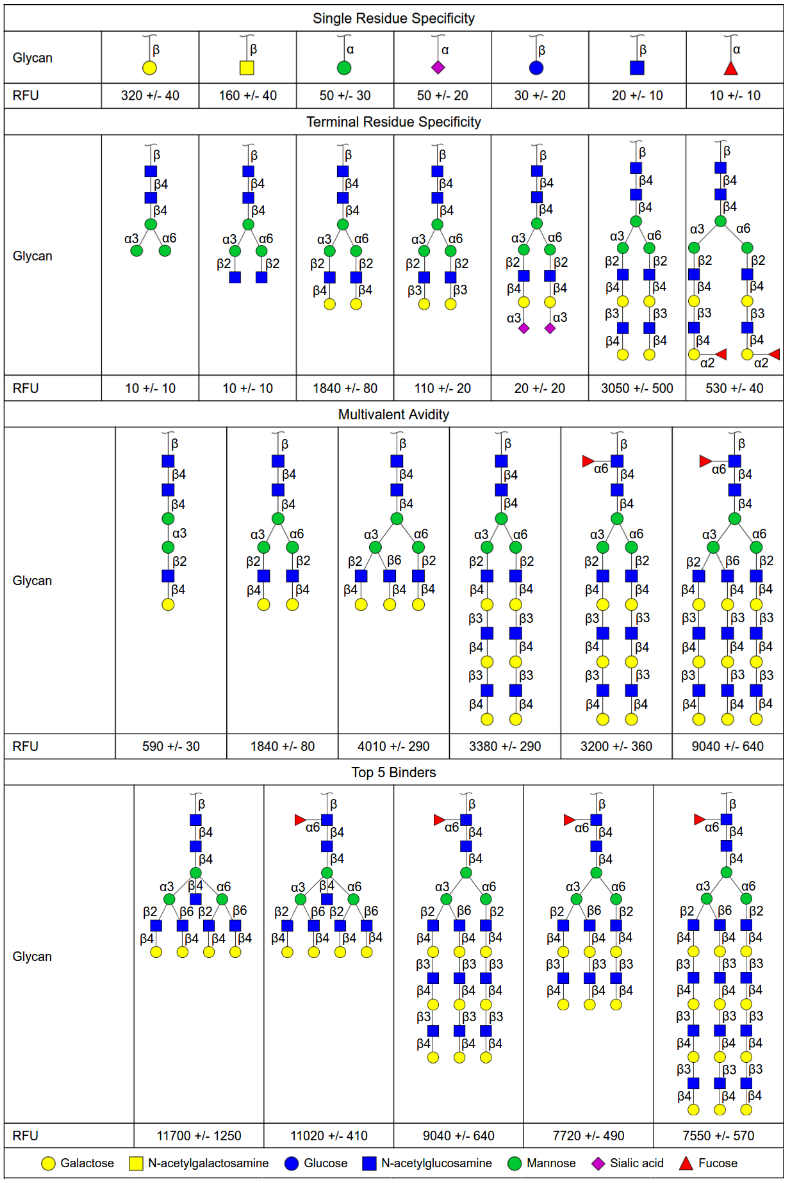


Figure 5.8: Structural analysis of ThreeFoil and OneFoil in the presence of lactose. (A) The change in relative fluorescence as a function of lactose concentration is well fit by a hyperbolic binding curve, giving a dissociation constant, K_d, of 1.22 ± 0.15 mM (see Methods). Similarly, galactose binding was measured previously to have a K_d of 1 mM [44]. No increase in fluorescence was seen with sucrose (data not shown), consistent with the lack of interaction in kinetic experiments (Figure 5.3). A glycomics screen for ThreeFoil revealed a strong preference for binding of multivalent carbohydrates terminating in galactose; protein monosaccharide binding is often in the mM range, with multivalent binding commonly observed to greatly increase affinity [44]. (B) The circular dichroism (CD) spectrum of: native ThreeFoil (without carbohydrate - solid black line, 50 mM lactose - dashed dark blue line, 50 mM sucrose - dotted dark red line); denatured from inclusion bodies (6 M urea) ThreeFoil (without carbohydrate - solid grey line, 50 mM lactose - dashed light blue line, 50 mM sucrose - dotted light red line) and native OneFoil (solid green line). All samples were in 300 mM KF and 100 mM sodium phosphate, pH 6.6. (C) Structural analysis for native ThreeFoil (black bars) and OneFoil (green bars) showing helical, beta, turn and disordered content analyzed using CDSSTR [279]. The predicted number of beta-strands is also shown as a fraction of the total in the crystal structure (PDB: 3PG0).



Single Residue Specificity

Glycan							
RFU	320 +/- 40	160 +/- 40	50 +/- 30	50 +/- 20	30 +/- 20	20 +/- 10	10 +/- 10

Terminal Residue Specificity

Glycan							
RFU	10 +/- 10	10 +/- 10	1840 +/- 80	110 +/- 20	20 +/- 20	3050 +/- 500	530 +/- 40

Multivalent Avidity

Glycan						
RFU	590 +/- 30	1840 +/- 80	4010 +/- 290	3380 +/- 290	3200 +/- 360	9040 +/- 640

Top 5 Binders

Glycan					
RFU	11700 +/- 1250	11020 +/- 410	9040 +/- 640	7720 +/- 490	7550 +/- 570

Figure 5.9: Glycan binding by ThreeFoil was tested *via* a glycan array containing over 600 mammalian glycans. Here selected data demonstrate that ThreeFoil prefers binding galactose as the terminal residue, particularly in a β-1,4 linkage. Note that lactose (binding shown in Figure 5.8) is composed of galactose in a β-1,4 linkage to glucose. In addition to terminal residue and linkage specificity, binding is dramatically enhanced when glycans are multivalent (multivalent avidity). Together this demonstrates that ThreeFoil function as a highly specific glycan binding protein or lectin.

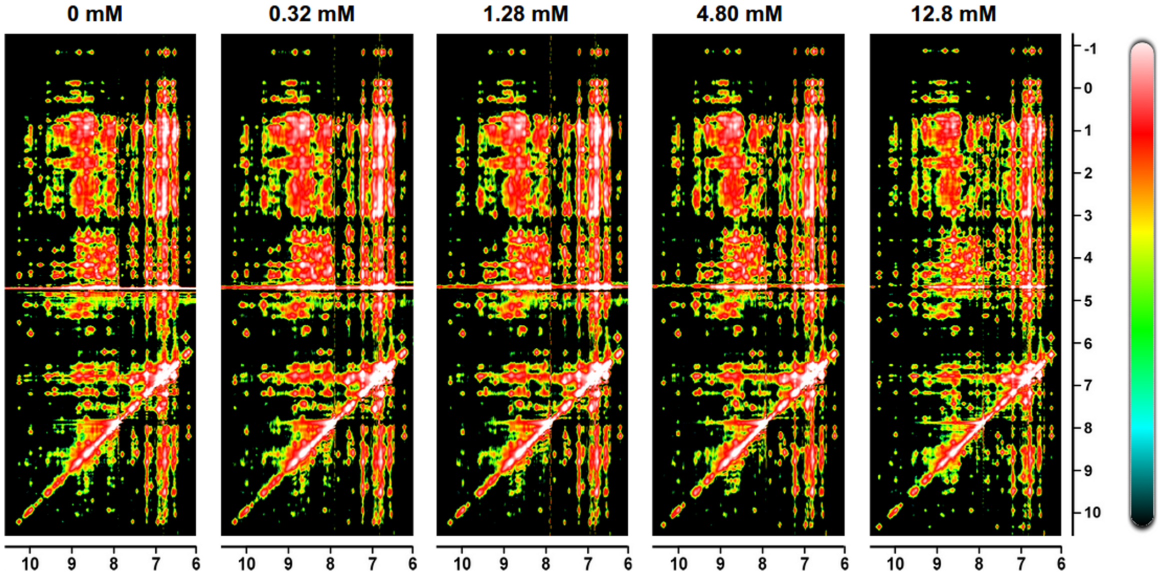


Figure 5.10: Lactose binding causes little change in the NOESY spectrum of ThreeFoil. 2D ^{1}H-^{1}H NOESY spectra shown in the amide region (∼6 ppm to ∼10.5 ppm) exhibit small shifts in the positions of a small number of resonances upon titration of the protein with lactose (from left to right: 0, 0.32, 1.28, 4.80, and 12.8 mM lactose). Color scale bar indicates the intensity of the observed peaks (black - no intensity, white - highest intensity). That the majority of peaks show no changes, combined with line widths remaining unchanged upon addition of lactose, indicates that the structural changes from lactose binding are local and do not account for the substantial increase in folding *m*-value and hence change in solvent accessible surface area observed for folding upon addition of lactose (Figure 5.3, Table 5.1).

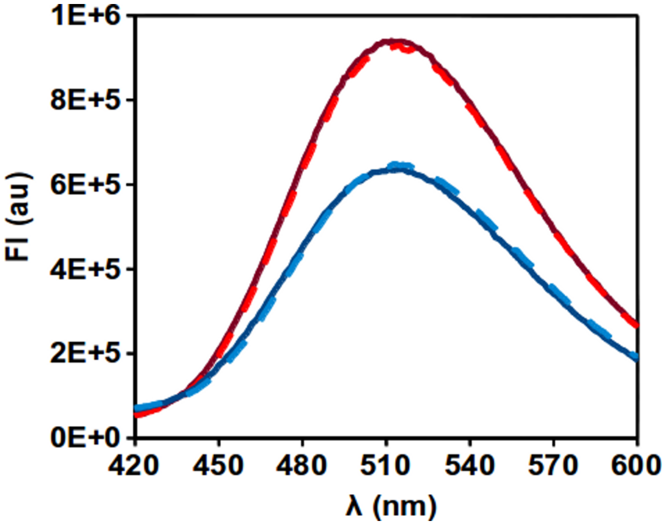


Figure 5.11: Native and denatured ThreeFoil do not bind ANS. Fluorescence (arbitrary units) of 8-anilinonaphthalene-1-sulfonic acid (ANS, 150 µM) in standard buffer (100 mM sodium phosphate, pH 6.6, with 300 mM NaCl) in the absence (dashed lines) or presence of ThreeFoil (1 µM, solid lines), was measured under native (blue lines) or denaturing conditions (6 M urea, red lines). ANS binding to clusters of exposed hydrophobic residues is characterized by marked increases in fluorescence intensity and a shift of the spectrum to lower wavelengths [280]; the absence of these spectral changes for ThreeFoil indicates a lack of ANS binding.

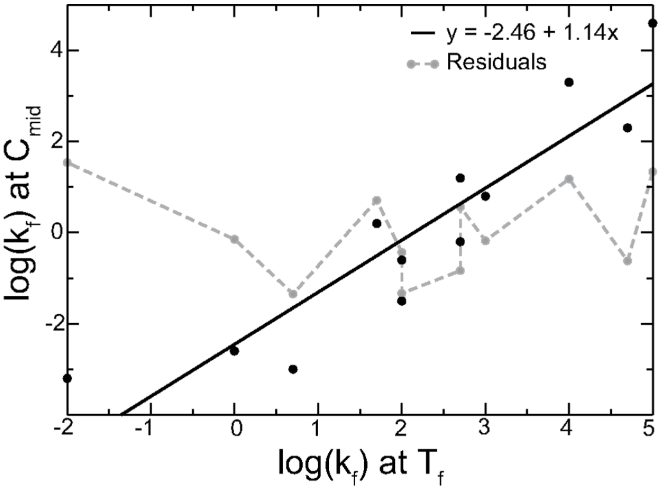


Figure 5.12: Experimental protein folding rates at midpoints of chemical and thermal denaturation are strongly correlated. $\log(k_f)$ at experimental denaturation midpoint, C_{mid}, plotted *versus* $\log(k_f)$ at the modelled folding temperature, T_f (black circles), where the folding rate equals the unfolding rate (k_f=k_u). The C_{mid} and T_f data are extracted from work by Broom and co-workers [78] and Chavez and co-workers [226], respectively. The black line gives the linear regression fit to these data (R = 0.91). The grey dashed line shows the residuals from the fit, which are distributed about 0. Proteins used in this data, from lowest "$\log(k_f)$ at T_f" to highest: Muscle acylphosphatase, 18th module of muscle protein twitchin, Titin IG repeat 27, Src SH3 domain, Colicin E9 immunity protein, B1 domain of Protein L, Chymotrypsin inhibitor 2, B1 domain of protein G, Cold shock protein, *Bacillus subtilis*, N-terminal domain of protein L9, Pyruvate dehydrogenase E2, *Pyrobaculum aerophilum*, Lambda repressor, and Villin headpiece.

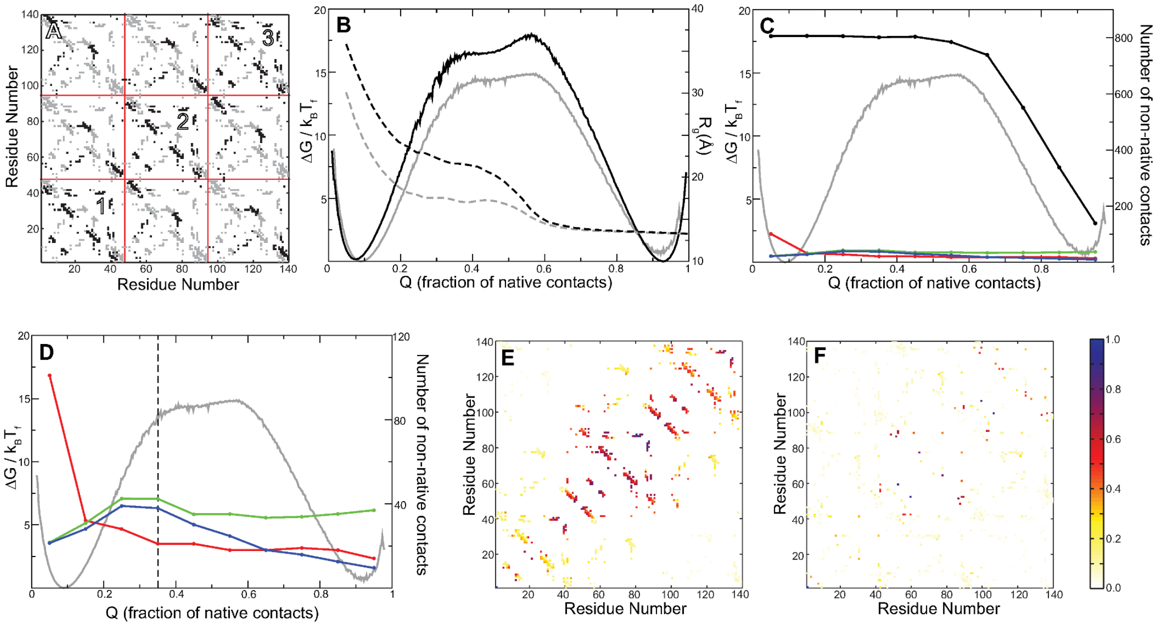


Figure 5.13: Simulations show symmetric non-native interactions do not trap ThreeFoil folding. (A) Symmetric non-native contacts in ThreeFoil. A black square at (x,y) means that residue x and y are in contact in the crystal structure of native ThreeFoil. The grey (non-native) contacts are obtained by calculating all contacts between pairs of residues that are symmetrically equivalent to residues x and y. The numbered red squares contain only intra-repeat contacts. (B) The folding free energies of ThreeFoil without (black) and with non-native contacts (grey) in units of their respective folding temperatures (k_BT_f, left Y axis) are plotted at (T_f) as a function of the fraction of native contacts (X axis). R_g of the two proteins (dashed lines) are plotted in the corresponding colors (right Y axis). The protein with the non-native interactions has a lower R_g in the unfolded state. (C) The number of non-native contacts (right Y axis) formed (contacts with Q_{nn}>0) is plotted in black as a function of native contacts (X axis). The number of non-native contacts whose probability of contact formation (Q_{nn}) is greater than the average native Q (X axis) is plotted in red. The average probability of contact formation of a non-native contact multiplied by the total number of non-native contacts is plotted as a function of Q in green. The non-native contact map from the model with stabilized non-native interactions is compared with the model with only native contacts by calculating the $RMSD$ of the two contact maps. Plotted in blue as a function of Q. It can be seen that the red, blue and green lines are close to 0. The free energy profile of the protein with non-native interactions is shown in grey (left Y axis) for reference. (D) Enlarged version of C to show features of the red, blue and green lines. The green (average probability of contact formation) and the blue ($RMSD$ from the native only model) lines are similar before the transition state (dotted line) meaning the protein with the non-native interactions makes more non-native contacts than the protein with only native contacts stabilized. However, post transition state, many of the non-native contacts that are formed are commensurate with the native state (compare E and F) and are also present in the native-only model simulations. This is why the value of the $RMSD$ reduces past the transition state. The scale of the free energy profile (grey; left Y axis) of the protein is the same as in C. (E) Average native contact map (A, black contacts) of the symmetrized non-native contact model at Q ∼0.35 (dotted line in D). Contacts are colored based on how formed they are. (F) Average non-native contact map (A; grey contacts) of the same model as E at Q ∼0.35. More formed contacts are close to the native contacts.

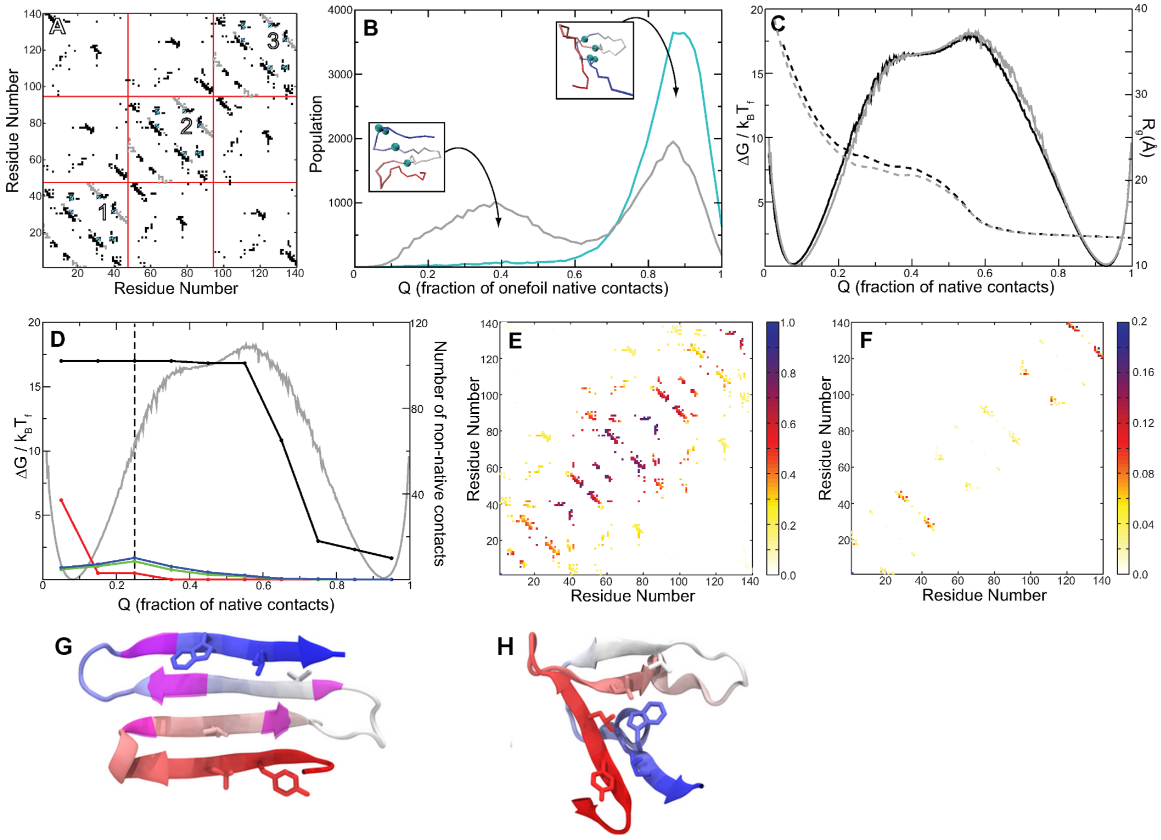


Figure 5.14: Coarse-grained simulations show that ligand binding reduces Rosetta-based non-native structure but leaves ThreeFoil folding largely unchanged

(A) Rosetta derived non-native contacts for ThreeFoil are shown in grey. These contacts are derived from the most populated structure of OneFoil from Rosetta simulations (see panel G), and are intra-trefoil (entirely within the numbered red squares). Black contacts are calculated from the crystal structure of ThreeFoil. The cyan contacts are between residues in the sugar binding site (Figure 5.2B). (B) Simulations of OneFoil (with both native and non-native contacts) are used to determine the strength of the non-native contacts. A strength of 1.5 times the strength of the native contacts was chosen because those simulations gave an almost equal population of native-like OneFoil (right inset) and Rosetta-like OneFoil (left inset) structures (grey curve). The lactose binding residues (cyan spheres; insets) are proximal in the native OneFoil but separated in Rosetta OneFoil. Four native contacts (cyan in A) are present between these residues. When their strength is doubled (representative of lactose binding) the population of the non-native Rosetta OneFoil is drastically reduced (cyan curve). (C) The folding free energies of ThreeFoil (black) and ThreeFoil with Rosetta non-native contacts (grey) in units of their respective folding temperatures ($k_B T_f$, left Y axis) are plotted at the transition midpoint (T_f) as a function of the fraction of native contacts (X axis). The radius of gyration of the two proteins (dashed lines) are plotted in the corresponding colors (right Y axis). There is limited change in either free energy or radius of gyration between the two conditions. (D) The number of formed non-native contacts (contacts for which Q_{nn}>0; right Y axis) is plotted in black as a function of native contacts (X axis). The number of non-native contacts whose probability of contact formation (Q_{nn}) is greater than the average native Q (X axis) is plotted in red. The average probability of contact formation of a non-native contact multiplied by the total number of non-native contacts is plotted as a function of Q in green. The non-native contact map from the model with stabilized non-native interactions is compared with the non-native contact map from the model with only native contacts by calculating the root mean square deviation of the two contact maps. This is plotted in blue as a function of Q. It can be seen that all the lines drop to 0 soon after the transition state (dotted line; Q ∼0.35). The green and the blue lines are similar throughout. This is because the non-native contacts are chosen such that they are not compatible with the WT ThreeFoil structure. The free energy profile of the protein with non-native interactions is shown in grey (left Y axis) for reference. (E) Average native contact map (A; only black and cyan contacts) of the protein with non-native contacts plotted at Q ∼0.35 (dotted line in D). Contacts are colored based on how formed they are and the color scale is given on the right. (F) Average non-native contact map (A; grey contacts) of the same model as E plotted at Q ∼0.35. Note that the color scales on E and F are different. A given non-native contact is formed to a significant extent only when the repeat (OneFoil) that it is a part of is largely unstructured. (G) The structure of OneFoil (a single repeat of ThreeFoil) for the lowest energy structure predicted by Rosetta *ab initio* [260] compared with the structure of a single repeat in the ThreeFoil crystal structure shown in H (PDB: 3PG0). The ribbon is coloured from red to white to blue from the N- to the C-terminus, with side-chains from the ThreeFoil hydrophobic core shown as sticks. Regions of additional beta-structure in the prediction compared with the crystal structure are shown in magenta. Similar structures to the one shown at left were adopted in $>\frac{1}{3}$ of the Rosetta predictions, while the native structure shown at right was very rarely predicted. Significant beta structure in OneFoil is also observed by CD (Figure 5.8B,C).

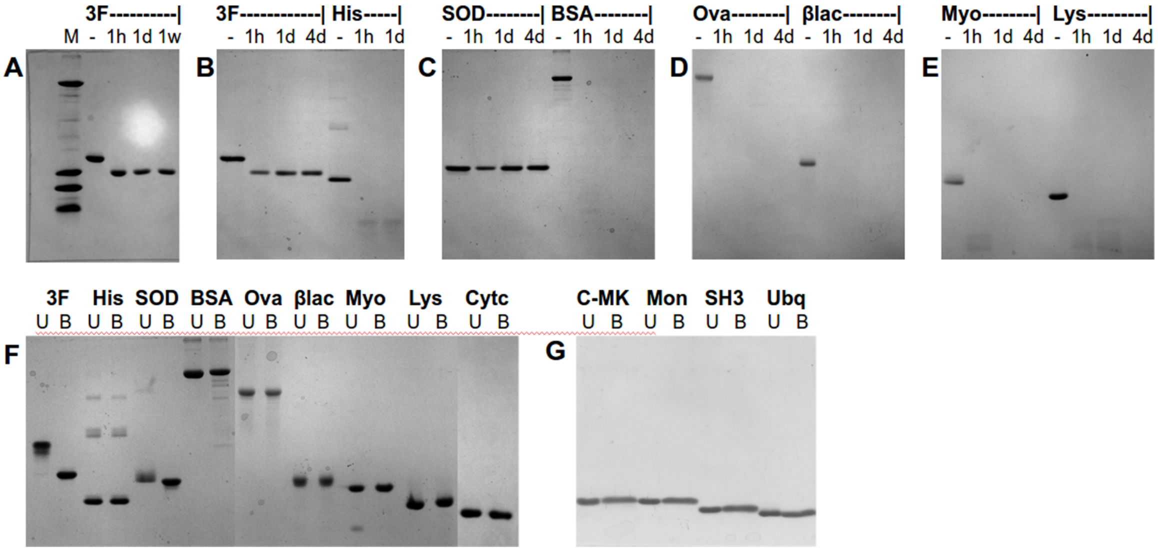


Figure 5.15: Detergent and protease resistance of ThreeFoil and other proteins. (A) The reduction in molecular weight for ThreeFoil (1 mg/mL) when incubated in Proteinase K (0.02 mg/mL) for up to a week results from removal of the cleavable-his tag. The molecular weight marker (lane M) includes BSA (66 kDa), myoglobin (17.6 kDa), hen egg-white lysozyme (14.3 kDa) and aprotinin (6.5 kDa). Tagged-ThreeFoil prior to incubation (lane C) and after 1 hour (1h), 1 day (1d) and 1 week (1w) show that the apparent molecular weight of ThreeFoil after incubation is 15.4 kDa in excellent agreement with the calculated molecular weight of the untagged protein at 15.3 kDa. Individual time-courses of protease resistance challenge are shown in panels B-E proteins (0.5 mg/mL) were incubated with Proteinase K (0.02 mg/mL), a nonspecific protease. Protease digestion SDS-PAGE analyses for (B) ThreeFoil (3F) and hisactophilin (His), (C) human Cu,Zn superoxide dismutase (SOD) and bovine serum albumin (BSA), (D) ovalbumin (Ova) and β-lactoglobulin (βlac), and (E) myoglobin (Myo) and lysozyme (Lys) are shown. Time points were taken: prior to addition of protease (-), 1 hour (1h), 1 day (1d), 4 days (4d). For simplicity Figure 5.6A shows only the pre-addition control (-) and the time when complete digestion was observed (1 hour for all but ThreeFoil and SOD) or the final time of the challenge (4 days) when no digestion was observed (ThreeFoil and SOD). SDS resistance tests with reducing agent are shown in panels F and G. Panel F shows the same test as in Figure 5.6B, but with β-mercaptoethanol added as reducing agent (7% v/v). Proteins (0.50 mg/mL) in SDS (1% w/v) were either unboiled (U) or boiled (B) for 10 minutes. From left to right: ThreeFoil (3F), hisactophilin (His), human Cu,Zn superoxide dismutase (SOD), bovine serum albumin (BSA), ovalbumin (Ova), β-lactoglobulin (βlac), myoglobin (Myo), lysozyme (Lys) and Cytochrome c (Cytc, not included in Figure 5.6B). Panel G is an identical test performed with different equipment and demonstrating a lack of resistance for the C-terminal domain of MK0293 (C-MK), single chain monellin (Mon), PI3K SH3 domain (SH3) and ubiquitin (Ubq). Where the unboiled sample runs the same as the boiled sample, no resistance to SDS is observed, whereas observation of a band at a higher position for the unboiled sample indicates that SDS is unable to penetrate and bind to the structure without the assistance of thermal unfolding. SOD is SDS-resistant in the absence of reducing agent, but loses resistance in its presence due to the reduction of its conserved disulfide bond. Threefoil does not contain disulfide bonds and remains SDS-resistant under these conditions. All data included in Table 5.3.

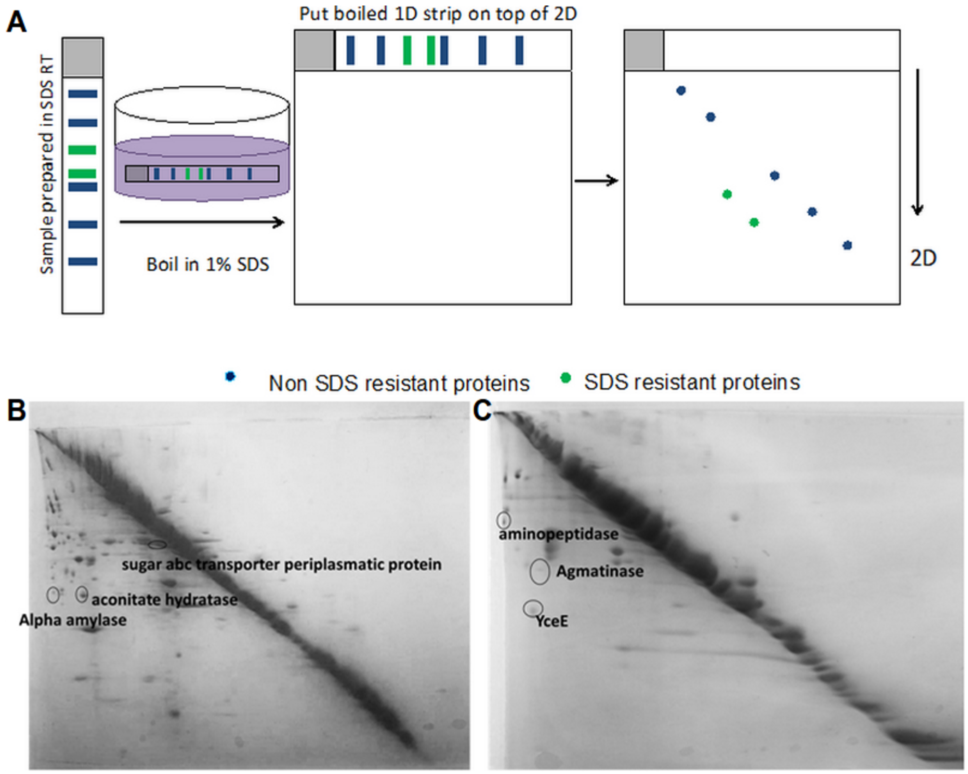

A
Sample prepared in SDS RT
Boil in 1% SDS
Put boiled 1D strip on top of 2D
2D
● Non SDS resistant proteins
● SDS resistant proteins
B
sugar abc transporter periplasmatic protein
aconitate hydratase
Alpha amylase
C
aminopeptidase
Agmatinase
YceE

Figure 5.16: Application of D2D SDS-PAGE for identifying KSPs at a proteomics level. (A) Diagram illustrating D2D SDS-PAGE. Unheated protein samples are analyzed by SDS-PAGE, and the relevant gel strip is then excised and heated in a boiling buffer solution containing SDS. The heated gel strip is analyzed by a 2nd dimension SDS-PAGE. The resulting 2D gel exhibits a diagonal pattern from the SDS-sensitive proteins migrating the same distance in both SDS-PAGE runs. However, SDS-resistant proteins migrate less in the 1st dimension, and thus end up below the gel diagonal. (B) D2D SDS-PAGE analysis of Thermus thermophilus. Protein spots below the gel diagonal bands were picked and analyzed by MALDI TOF. The MS data was analyzed by MASCOT Peptide Mass Fingerprint (PMF) search against the whole NCBInr database without specifying taxonomy. Only significant protein Identification with $p<0.05$ and taxonomy from *T. thermophilus* were accepted. Circled spots identify the monomeric (single domain with no disulfide bonds) KSPs identified. (C) D2D SDS-PAGE analysis of *Bacillus subtilis*. Protein spots below the gel diagonal bands were picked and analyzed by LC/MS/MS (LTQ-Orbitrap). Mass spectrometry data obtained from all LC-MS-MS analysis were searched against Swissprot using Sequest search algorithms through Proteome Discoverer. Commonly accepted criteria for high-confidence peptide identifications (xCorr 1.8 for +1, 2.5 for +2, 3.5 for +3) was used to screen peptides. To achieve low false positive rates of protein identification, each protein included in the results table contains more than 2 high-confidence unique peptides. Circle spots identify the monomeric (single domain with no disulfide bonds) KSPs identified which had known structures (or structures for highly homologous sequences were known).

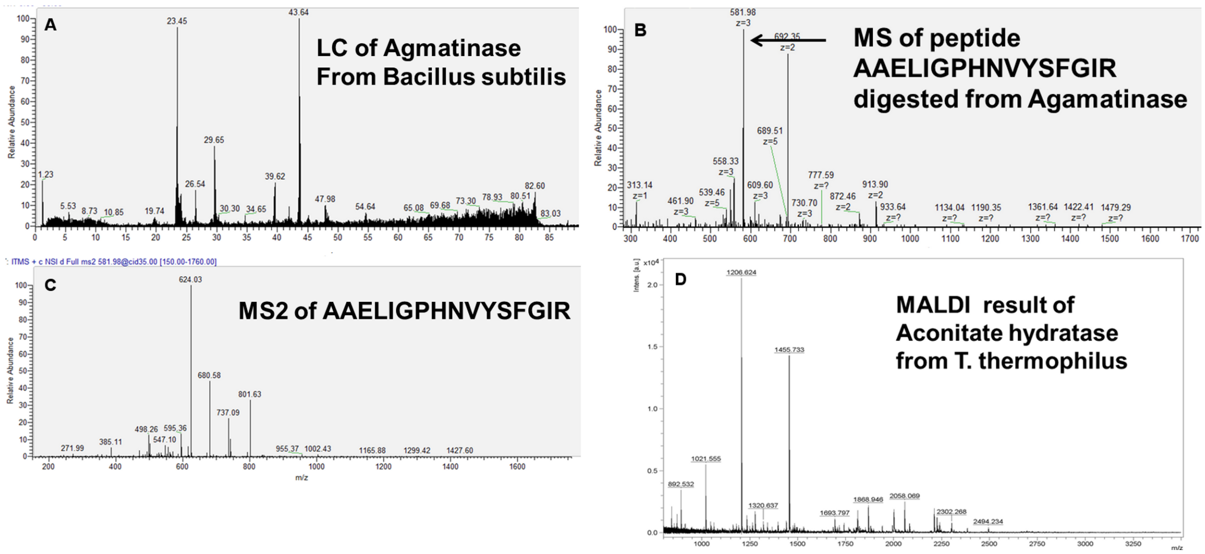


Figure 5.17: Examples of peptide identification by mass spectrometry. Top 3 panels are LC-MS-MS analysis of Agmatinase from *Bacillus subtilis* as example of *Bacillus subtilis* KSP identification. Mass spectrometry data obtained from all LC-MS-MS analysis were searched against Swissprot using Sequest search algorithms through Proteome Discoverer. Commonly accepted criteria for high-confidence peptide identifications (xCorr 1.8 for +1, 2.5 for +2, 3.5 for +3) was used to screen peptides. To achieve low false positive rates of protein identification, each protein included in the results table contains more than 2 high-confidence unique peptides. Agmatinase identification contains 9 high-confidence unique peptides including AAELIGPHNVYSFGIR. On the bottom is MALDI result of Aconitate hydratase from *Thermus thermophilus* as example of *Thermus thermophilus* KSP identification. The MS data was analyzed by MASCOT Peptide Mass Fingerprint (PMF) search against the whole NCBInr database without specifying taxonomy. Only significant protein Identification with p<0.05 and taxonomy from *Thermus thermophilus* were accepted. For Aconitate hydratase, 11 peptides matched, MASCOT score 109 and expectation value is 8.9x10^{-4} which is significantly below 0.05.

Table 5.1: Folding and unfolding kinetics of ThreeFoil, Symfoil, and Hisactophilin

Protein	$\ln(k_f^{H2O})$	$\ln(k_u^{H2O})$	m_f (kcal mol^{-1} M^{-1})	m_u (kcal mol^{-1} M^{-1})	m (kcal mol^{-1} M^{-1})	ΔG_{U-F}^{H2O} (kcal mol^{-1})
ThreeFoil - sodium	-8.98	-15.0	-7.8	3.19	11.0	3.6
ThreeFoil	-9.15	-19.7	-6.8	2.94	9.7	6.3
ThreeFoil + sucrose	-9.06	-20.0	-6.8	2.98	9.8.0	6.5
ThreeFoil + lactose	-7.44	-29.8	-7.7	4.29	12.0	13.3
Symfoil (7)	4.9	-13.8	-7.9[a]	1.5[a]	9.4[a]	11.1
Hisactophilin (8)	3.1	-10.8	-8.8[a]	5.1[a]	13.9[a]	8.2

The expected value for complete unfolding of ThreeFoil is 14.2 kcal mol^{-1} M^{-1} (see Materials and Methods)

[a] Values measured in GuHCl for Symfoil [79] or urea for Hisactophilin [249] are estimated in GuSCN for comparison (see Methods)

Table 5.2: Proteins with experimentally determined kinetics

Name	PDB	Length[a]	ACO	LRO	Structure	$\ln(k_u) = \ln(k_f)$
Colicin E7 immunity protein	1AYI	85	9.3	2.54	alpha	2.8
Telomeric protein DNA-binding domain, human	1BA5	49	6.8	2.20	alpha	1.6
Immunoglobulin binding B-domain	1BDD (2-59)	58	5.8	2.17	alpha	5.8
16th domain of brain alpha-spectrin	1CUN (7-112)	106	11.6	2.34	alpha	-2.0
17th domain of brain alpha-spectrin	1CUN (113-219)	107	12.0	2.56	alpha	-3.4
FADD death-domain, human	1E41 (93-192)	100	8.2	3.02	alpha	-0.6
Rap1 myb-domain, human	1FEX	59	6.8	2.27	alpha	3.9
Myb transforming protein	1IDY	54	6.0	1.59	alpha	3.1
Colicin E9 immunity protein	1IMQ	85	11.3	3.51	alpha	-1.4
Trp-Cage Miniprotein	1L2Y	20	4.1	0.50	alpha	13.0
Lyme disease variable surface antigen	1L8W (29-335)	307	25.0	5.30	alpha	-4.7
Lambda repressor	1LMB	80	8.2	2.63	alpha	5.2
Acyl-coenzyme A binding protein, cow	1NTI	86	11.2	3.56	alpha	-0.3
Protein yjbJ	1RYK	69	7.9	2.72	alpha	6.4
BBA5 mini-protein	1T8J	23	3.2	0.35	alpha	12.5
15th domain of brain alpha-spectrin	1U5P	110	12.4	2.60	alpha	4.0
Villin headpiece	1VII	36	4.3	0.89	alpha	10.6
Dihydrolipolysine acetyltransferase, *G. stearothermophilus*	1W4G	45	6.2	2.36	alpha	5.8
Dihydrolipolysine succinyltransferase, *E. coli*	1W4H	45	5.5	2.00	alpha	9.2

Pyruvate dehydrogenase E2, *P. aerophilum*	1W4J	51	5.8	2.31	alpha	7.5
de novo designed 3-helix bundle	2A3D	73	7.4	2.14	alpha	12.4
Peripheral subunit-binding domain, dihydrolipoamide acetyltransferase	2PDD	42	5.4	2.05	alpha	9.8
E3-binding domain of BBL	2WXC	47	4.9	1.74	alpha	8.0
Cold shock protein, *B. caldolyticus*	1C9O	66	12.0	4.42	beta	0.2
Cold shock protein, *B. subtilis*	1CSP	67	12.1	4.69	beta	2.7
Forming-binding protein 28	1E0L	37	6.7	2.16	beta	9.2
WW prototype	1E0M	37	6.3	2.11	beta	7.7
9th fibronectin domain	1FNF	90	17.3	5.51	beta	-2.2
Cold shock protein, *T. maritima*	1G6P	66	12.8	5.24	beta	-2.6
Hisactophilin	1HCE	117	13.6	4.29	beta	-6.5
sso7d	1JIC	62	6.7	2.52	beta	0.6
Abp1 SH3 domain	1JO8	58	12.1	4.93	beta	-2.1
FGF-1	1JQZ	136	17.8	5.82	beta	-6.6
Fibronectin type III WL-12 chitinase A1	1K85 (559-644)	86	15.6	5.14	beta	-3.8
E2 component alpha-ketoacid dehydrogenase	1K8M	87	17.9	5.33	beta	-4.7
Yes kinase-associated protein	1K9Q (5-44)	40	7.2	2.70	beta	7.0
Internalin B SH3 domain	1M9S (391-466)	76	14.8	4.61	beta	0.1
Cold shock protein, *E. coli*	1MJC	69	11.9	4.55	beta	1.6
beta-hairpin	1PGB (41-56)	16	4.5	0.63	beta	12.0
PinWW	1PIN (6-36)	34	7.0	2.24	beta	9.1
PI3K SH3 domain	1PNJ	84	15.5	4.74	beta	-4.8

Oncoprotein p13mtcp1	1QTU (1-109)	109	17.0	5.17	beta	-8.5
Fyn SH3 domain	1SHF	59	11.9	4.75	beta	-2.0
Spectrin SH3 domain	1SHG	57	11.9	5.05	beta	-3.7
Src SH3 domain	1SRL	56	11.8	4.39	beta	0.4
Fibronectin type III tenascin	1TEN	89	16.3	5.33	beta	-4.4
18th module of muscle protein twitchin	1WIT	93	20.2	6.19	beta	-5.9
Sho1 SH3 domain	2VKN	66	14.3	4.79	beta	-0.8
Symfoil1	3O49	123	15.8	5.43	beta	-7.1
Symfoil4P	3O4D	123	16.3	5.30	beta	-10.8
ThreeFoil	3PG0	140	23.0	6.31	beta	-17.1
Muscle acylphosphatase	1APS	98	22.2	5.78	mixed	-7.3
C-terminal domain of protein L9	1DIV (58-149)	92	13.9	4.17	mixed	-4.4
N-terminal domain of protein L9	1DIV (1-56)	56	7.8	2.86	mixed	1.9
LysM domain	1E0G	48	10.1	3.29	mixed	3.7
FK506 binding protein	1FKB	107	20.1	5.42	mixed	-5.2
Apoflavodoxin, Anabaena	1FTG	168	17.5	6.07	mixed	-0.2
Tm1083	1J5U	123	20.9	5.23	mixed	-0.9
Chemotaxis protein CheW	1K0S	143	20.9	5.38	mixed	-5.0
Cyclophilin A	1LOP	164	27.9	5.99	mixed	-3.0
Ribosomal protein L23	1N88	96	14.6	5.02	mixed	-2.4
ADAh2	1O6X	81	12.8	3.53	mixed	1.5
B1 domain of protein G	1PGB	56	9.8	3.21	mixed	-0.4
Histidine containing phosphocarrier protein	1POH	85	15.9	4.56	mixed	-3.0
C-terminal domain of spore coat protein S	1PRS (91-173)	83	14.5	5.76	mixed	-8.1
N-terminal domain of spore coat protein S	1PRS (1-90)	90	14.8	5.69	mixed	-4.6
Ras binding domain	1RFA	78	13.8	4.64	mixed	-0.2

Ribosomal protein S6	1RIS	97	20.0	4.93	mixed	-3.9
Src SH2 domain	1SPR	103	12.9	4.08	mixed	-1.4
Ubiquitin	1UBQ	76	12.6	4.18	mixed	-1.5
Spliceosomal protein U1A	1URN	96	17.9	4.69	mixed	-0.4
Common-type acylphosphatase	2ACY	98	21.4	5.78	mixed	-4.4
Chymotrypsin inhibitor 2	2CI2	64	10.8	4.28	mixed	-3.5
B1 domain of Protein L	2PTL (18-77)	60	12.0	4.10	mixed	-1.4
Pit1 homeodomain	1AU7 (103-160)	58	5.8	1.69	alpha	7.4
FKBP-Rapamycin binding domain	1AUE	92	10.7	3.09	alpha	-2.4
p19ink4d CDK inhibitor	1BD8	156	8.7	3.90	alpha	-1.9
Engrailed homeodomain	1ENH	54	7.9	2.15	alpha	8.1
Acyl-coenzyme A binding protein, yeast	1ST7	86	11.3	3.14	alpha	-0.4
FF domain, human HYPA	1UZC	69	9.4	3.51	alpha	2.3
Tumor suppressor protein p16	2A5E (9-156)	148	9.3	4.22	alpha	0.4
Phage 434 cro protein	2CRO	64	8.0	2.41	alpha	0.3
T4 lysozyme	2LZM	164	12.4	3.27	alpha	-5.7
Myotrophin	2MYO	118	7.7	3.53	alpha	-0.3
Ileal lipid binding protein	1EAL	127	17.2	4.41	beta	-2.8
Intestinal fatty acid binding protein	1IFC	131	18.9	4.75	beta	-2.8
Cellular retinol binding protein II	1OPA	133	19.8	4.89	beta	-5.0
Barnase	1RNB	109	12.3	4.31	beta	-4.3
Titin IG repeat 27	1TIU	89	17.2	5.33	beta	-6.9
10th type III fibronectin domain	1TTF	94	12.0	4.06	beta	-0.5
Apical domain of GroEL	1AON (191-345)	155	23.0	5.21	mixed	-3.4

PSD-95, third PDZ domain	1BFE	110	17.3	4.96	mixed	-0.2
Barstar	1BTA	89	11.7	4.63	mixed	-1.4
Cellular retinoic acid binding protein I	1CBI	136	19.6	4.62	mixed	-6.7
PDZ2 domain from PTP-BL	1GM1	94	17.4	4.94	mixed	-1.2
Hydrogenase maturation protein	1GXT	88	19.7	5.43	mixed	-0.8
Indole-3-glycerolphosphate synthase	1IGS (27-248)	222	20.2	5.79	mixed	-8.9
Staphylococcal nuclease	1JOO	149	13.3	4.21	mixed	-5.3
C-terminal domain of phosphoglycerate kinase	1PHP (176-394)	219	19.1	5.54	mixed	-5.1
N-terminal domain of phosphoglycerate kinase	1PHP (1-175)	175	21.9	5.62	mixed	-1.3
Trp-Synthase alpha-subunit	1QOP (Chain A)	265	23.9	5.63	mixed	-5.7
Dihydrofolate reductase	1RA9	159	24.5	5.50	mixed	-5.2
Cell-cycle regulatory protein p13suc1	1SCE	97	8.1	2.37	mixed	-1.0
Carbonic anhydrase	5BNL	257	38.3	6.89	mixed	-10.6
Ribonuclease H1, *E. coli*	2RNR	155	21.0	5.30	mixed	-4.6
Villin 14T	2VIK	126	16.9	5.40	mixed	-1.6
Chemotactic protein	3CHY	128	12.0	4.56	mixed	-1.3
Ribonuclease H1, *C. tepidum*	3H08	139	11.5	3.96	mixed	-4.0

[a] Number of structured residues (from PDB) in the domain

Table 5.3: **Resistance dataset of monomeric proteins**

Name	PDB	Length[a]	ACO	LRO	Structure	Source
Resistant, Disulfide-Free, Single Domains (Figure 5.6C,D "Resistant")						
ThreeFoil	3PG0	140	23.0	6.31	beta	this work
OmpX	1QJ8	148	20.9	6.49	beta	[272]
YceE	3IBZ	176	23.1	5.88	beta	this work
LicB lichenase	3WVJ	219	37.2	6.92	beta	[290]
Green fluorescent protein	1EMA	221	29.0	6.49	beta	[291]
Glutamine binding protein	1WDN	223	27.1	5.61	mixed	[271]
Endo-β-N-acetylglucosaminidase	1C3F	265	26.9	6.51	mixed	[292]
Agmatinase	3LHL	272	28.0	6.25	mixed	this work
Subtilisin Carlsberg	4C3U	274	31.0	7.07	mixed	[293]
Chitinase	3AFB	300	24.6	6.15	mixed	[294]
Phosphoenolpyruvate carboxykinase	1AYL	532	33.9	6.37	mixed	[271]
Not Resistant, Disulfide-Free, Single Domains (Figure 5.6C,D "Not Resistant")						
Rubredoxin	4RXN	54	10.4	3.33	limited	[245]
Ubiquitin	1UBQ	76	12.6	4.18	mixed	this work
C-terminal domain of MK0293	3C19[b]	79	10.4	3.80	mixed	this work
PI3K SH3 domain	1PNJ	86	15.5	4.74	beta	this work
Monellin single chain	1FA3	96	11.4	4.52	mixed	this work
Cytochrome c	1LC1	104	11.4	3.38	alpha	this work
Hisactophilin	1HCE	118	13.7	4.31	beta	this work
Myoglobin	1MCY	154	14.1	3.09	alpha	this work, [293]
Carbonic anhydrase	5BNL	257	38.3	6.89	mixed	[245]
Rhodanese	1DP2	281	28.6	6.22	mixed	[245]
Not Resistant, Disulfide-Linked, Single Domains (Figure 5.6C,D "Not Resistant")						
Insulin	1ZNI	21	3.4	0.67	limited	[245]

β-2-microglobulin	1BMG	98	17.5	5.73	beta	[245]
α-lactalbumin	2G4N	122	13.7	4.25	alpha	[245]
Ribonuclease A	5RSA	124	15.7	5.87	mixed	[245]
Hen egg-white Lysozyme	3WUN	129	14.9	4.43	alpha	this work
Trypsin	4MTB	223	29.0	6.97	beta	[245]
Resistant, Disulfide-Linked, Single Domains (not plotted)[c]						
Soybean trypsin inhibitor	1AVU	172	25.6	6.07	beta	[273]
Protease A	2OUA	188	25.1	7.45	beta	[295]
α-lytic protease	2ALP	198	25.4	7.19	beta	[295, 264]
Papain	1PPN	212	26.4	6.36	mixed	[245]
Chymopapain	1YAL	216	28.0	6.45	mixed	[245]
Proteinase K	2PRK	279	30.5	7.11	mixed	[296]
Putrescine binding protein	1A99	341	38.4	5.87	mixed	[271]
Resistant, Disulfide-Free, Multiple Domains (not plotted)[c]						
Ribose binding protein	2DRI	271	22.6	6.30	mixed	[271]
Aminopeptidase	1VHE	354	36.5	5.97	mixed	this work
Maltose binding protein	1ANF	369	39.6	5.68	mixed	[271]
Elongation factor Tu	1EFT	405	30.3	6.61	mixed	[272]
Sugar ABC transporter periplasmic protein	1EU8	407	42.6	5.96	mixed	this work
α-amylase catalytic subunit	4GKL	420	27.1	6.43	mixed	this work
Elongation factor EF2	1DAR	615	31.9	6.13	mixed	[272]
Aconitate hydratase	2B3X	888	52.6	6.65	mixed	this work
Resistant, Disulfide-Linked, Multiple Domains (not plotted[c]						
Pepsin B	3PEP	326	32.0	6.58	beta	[297]
UDP hydrolase	1HP1	516	36.7	6.64	mixed	[271]
Not Resistant, Disulfide-Linked, Multiple Domains (not plotted)						
γ-β-crystallin	4GCR	174	18.1	6.52	beta	[245]
Bovine serum albumin	4F5S	583	17.8	4.00	alpha	this work

Domains defined using “Domain Parser” [298] as in the RCSB (http://www.rcsb.org/). Compact domains are determined based on the density of residue-residue contacts

For “this work”, gels of resistance analysis are shown in Figures 5.6A,B, 5.15 and 5.16

Here we use SDS or protease resistance as evidence of a slowly unfolding structure with limited weak points. For resistance to protease, only proteases with broad specificity such as Proteinase B and Thermolysin are considered, whereas highly specific proteases such as Trypsin are not, as resistance in the latter case could simply result from few potential cut-sites

[a] Number of structured residues (from PDB) in the domain

[b] Residues 98-178 of the PDB coordinate file make up the experimentally tested domain

[c] Resistant, but disulfide-linked proteins are not plotted as some may not be resistant without cross-links. Nevertheless, they have similarly high ACO/LRO to those of resistant proteins without disulfides. Resistant multi-domain proteins also have a high ACO/LRO but are not plotted as topological complexity has not been confirmed to correlate with transition barrier heights beyond a single domain context (multimeric proteins are not included here for similar reasons). These data are included here for interested readers.

Chapter 6

Do Protein Stability Prediction Tools Work When Engineering Stabilizing Mutations?

6.1 Context

Determining the kinetics of ThreeFoil in the previous chapter showed it to have a fairly average thermodynamic stability (∼6 kcal/mol). I wanted to improve upon this so that future work looking at altering binding specificity would be possible. Since functional mutations tend to be destabilizing, I hoped to provide a large buffer for such mutations. Zachary Jacobi (a summer student) and I found a number of automated tools (either webservers or stand-alone programs) that would predict the thermodynamic effect of a point mutation. After predicting all possible point mutations to ThreeFoil, we chose to test the top 10 experimentally, which I grew, harvested, and purified. During Zachary's second term in the lab, he and I determined the kinetics of these ThreeFoil mutants and were disheartened by the limited improvements in stability we observed. We thus undertook a detailed analysis of a large number of prediction tools and the Protherm database they are universally tested on. The analysis revealed several flaws in the metrics that are commonly used to report the performance of these tools. I describe these flaws, how they might be overcome in the future, and what can be done now to get the best results, in this chapter.

6.2 Summary

Proteins often possess marginal stability. This limits production yield and shelf-life, while reducing the activity and usability of biocatalysts and increasing the likelihood of immunogenicity in biotherapeutics. Accurate computational tools to predict stabilizing point mutations, which can be fast and exhaustive, have been much sought after. Over the last decade a plethora of such tools have been reported. Accuracy is typically reported as ∼80% when tested against known mutations. However, later experimental application of these tools to stabilize proteins show poor success rates of ∼25%. Through a detailed analysis we find that many commonly reported performance metrics can be misleading. Possibly because of this, progress has been slow and uncertain. Additionally, datasets used for testing poorly reflect the mutations desired in biotechnology applications. To support the future development of effective stability prediction tools I provide guidelines for robust performance metrics and highlight the tools and general approaches giving the highest likelihood of successful stabilization at present.

6.3 Introduction

Improving the typically marginal stability of proteins [34, 33] is crucial to generating effective biocatalysts and biotherapeutics, yet is frequently a time-consuming and resource intensive undertaking. Biocatalysts can be used for many applications such as: chemical synthesis of valuable reagents or breakdown of biomass into biofuels, and increasing stability allows for elevated reaction temperatures which increase reaction rates and reduce unwanted bacterial growth [299]. In the case of biotherapeutics such as monoclonal antibodies, antibody mimetics [8, 300], and other binding scaffolds [65], increasing stability not only improves yield, but limits aggregation, improving shelf-life and reducing unwanted immune response [301, 302, 303]. For both biocatalysts and biotherapeutics, increasing stability improves the efficiency and effectiveness of experimental selection methods aimed at enhancing or modifying the original function [304, 153]. This functionalization is critical for moving beyond the natural repertoire of proteins or customizing designed scaffolds for unique applications. Improving stability has frequently been tackled using experimental selection methods such as directed evolution [299, 305]. While such methods are often successful, they require considerable resources and expertise. By contrast, automated computational tools require much less human involvement and are typically free to use. Such computational tools offer significant promise for expanding the range and effectiveness of protein biotechnology applications.

Evidence, however, continues to accumulate demonstrating this promise has yet to be realized. Using automated computation tools for predicting ΔΔG upon point mutation, we sought to improve the stability of our previously designed protein, ThreeFoil [44, 70]. After predicting all point mutations we experimentally tested the top 10 suggestions at diverse positions, finding only 2 were stabilizing. These poor results echo those seen by many other groups where the use of stability prediction tools to design mutations yielded stabilizing mutations in ∼25% of cases and even those were of underwhelming magnitude (Table 6.1, Supplementary Table 6.3). These results come as a great surprise given that reported accuracies for most of these tools are ∼80% when tested against the Protherm database (Table 6.1), a widely used and freely accessible database of protein point mutation data [306]. Though a previous smaller analysis of 6 tools found self-reported metrics were slightly inflated over independent testing, overall performance was fairly good [92]. It was, however, found that performance can vary greatly from one protein to another [92]. Is the disparity between the apparent success of these tools when tested against known data and their poor ability to reliably produce increases in stability when applied to new real-world problems, simply a matter of bad luck? Or, is there a fundamental problem with the way success is being evaluated? We find the latter to be the case.

Herein we report a detailed analysis of 21 popular tools (Table 6.2), and demonstrate that the aforementioned disparity between reported performance and experimental application originates from the frequent use of inappropriate performance metrics. Specifically, metrics such as accuracy and standard error poorly discriminate between predictive and nearly random tools and even Pearson's correlation coefficient (R) can be misleading given the need to remove outliers. By contrast, the Matthews and Spearman correlation coefficients (MCC and ρ) appropriately highlight the tools most capable of recommending stabilizing mutations, not only showing where future efforts in development may best be focused, but demonstrating the current best approaches for protein engineers hoping to stabilize their molecular machines.

Table 6.1: Success of stability prediction tools in designing stabilizing point mutations

Tool Used	# of mutations tested / # of reports[a]	Success Rate (PPV)	Average change in stability ($\Delta\Delta G_{exp,avg}$)
FoldX	49 / 4	14%	-0.42 ± 1.40
FoldX + MD	26 / 1	27%	0.0 ± 0.79
Rosetta-ddG	19 / 2	16%	-0.75 ± 1.87
Rosetta-ddG + MD	20 / 1	20%	-0.17 ± 0.83
PoPMuSiC	47 / 8	43%	0.53 ± 1.48
FoldX + TI	18 / 1	56%	0.66 ± 0.96
All	**148 / 12**[b]	**30%**	**0.03 ± 1.25**

For a complete mutation by mutation breakdown and source references, see Supplementary Table 6.3.

[a] # of reports refers to the number of publications in the literature that these mutations were pullled from.

[b] The total is less than the sum of the individual tools because some mutations were chosen by multiple tools (in cases where FoldX and Rosetta-ddG were used together for instance).

Table 6.2: Automated stability prediction tools

Tool / Publication Date	Method -level of detail -scoring -training	Typical run-time	Reported metrics	Robust metrics shown here
SDM 1997 [307]	-coarse-grained -statistical potential -simple scaling factor	Seconds	Acc: 86% *R*: 0.80 Error: N/A	MCC: 0.33 ρ: 0.47 $\Delta\Delta G_{exp,avg}$: -0.24
FoldX 2002 [308]	-atomistic -empirical potential -fitting (9 parameters to ΔΔG data	Minutes	Acc: N/A *R*: 0.73 Error: 1.0	MCC: 0.41 ρ: 0.55 $\Delta\Delta G_{exp,avg}$: 0.05
IMutant2 2005 [309]	-atomistic and coarse-grained -features -machine learning (43 inputs trained on ΔΔG data)	Seconds	Acc: 80% *R*: 0.71 Error: 1.3	MCC: 0.09 ρ: 0.31 $\Delta\Delta G_{exp,avg}$: -0.19
EGAD 2005 [310]	-atomistic -physical potential -fitting (3 parameters to ΔΔG data)	Minutes	Acc: N/A *R*: 0.69 Error: 1.0	MCC: 0.35 ρ: 0.54 $\Delta\Delta G_{exp,avg}$: 0.13
CUPSAT 2006 [311]	-atomistic -statistical potential -fitting (41 parameters to ΔΔG data)	Seconds	Acc: 80% *R*: 0.77 Error: 1.2	MCC: 0.22 ρ: 0.42 $\Delta\Delta G_{exp,avg}$: -0.11
MUpro 2006 [312]	-coarse-grained -features -machine learning (160 inputs trained on ΔΔG data)	Seconds	Acc: 85% *R*: 0.60 Error: 1.1	MCC: 0.13 ρ: 0.39 $\Delta\Delta G_{exp,avg}$: -0.21
Eris 2007 [313]	-atomistic -statistical potential -fitting (20 parameters to structural data only)	Hours	Acc: N/A *R*: 0.66 Error: 2.4	MCC: 0.32 ρ: 0.38 $\Delta\Delta G_{exp,avg}$: -0.15
MultiMutate 2007 [314]	-coarse-grained -statistical potential -none	Seconds	Acc: 80% *R*: N/A Error: N/A	MCC: 0.16 ρ: 0.40 $\Delta\Delta G_{exp,avg}$: -0.41
IMutant3 2008 [315]	-atomistic and coarse-grained -features -machine learning (43 inputs trained on ΔΔG data)	Seconds	Acc: 84% *R*: 0.69 Error: N/A	MCC: 0.13 ρ: 0.42 $\Delta\Delta G_{exp,avg}$: -0.46

DFire2 2008 [316]	-atomistic -statistical potential -simple scaling factor	Seconds	Acc: N/A R: N/A Error: N/A	MCC: 0.44 ρ: 0.57 $\Delta\Delta G_{exp,avg}$: 0.01
CC/PBSA 2009 [317]	-atomistic -physical potential -fitting (4 parameters to ΔΔG data)	N/A[a]	Acc: N/A R: 0.75 Error: 1.0.X	MCC: 0.40 ρ: 0.69 $\Delta\Delta G_{exp,avg}$: -0.18
PoPMuSiC2.0 2009 [318]	-coarse-grained -statistical potential and features -fitting *via* machine learning (27 parameters to ΔΔG data)	Seconds	Acc: N/A R: 0.63 Error: 1.2	MCC: 0.43 ρ: 0.60 $\Delta\Delta G_{exp,avg}$: 0.24
Hunter 2009 [92, 319]	-atomistic -statistical potential -fitting (4 parameters to ΔΔG data)	Seconds	Acc: 69% R: 0.45 Error: 1.09 (AUE)	MCC: 0.17 ρ: 0.34 $\Delta\Delta G_{exp,avg}$: -0.39
Thermodynamic Integration (TI)[b] 2010 [320]	-atomistic -physical potential -none	Days (estimated)	Acc: 88% R: 0.86 Error: 0.8 (AUE)	MCC: 0.43 ρ: 0.69 $\Delta\Delta G_{exp,avg}$: -0.22
Rosetta-ddG 2011 [321]	-atomistic -statistical and empirical -fitting (Rosetta forcefield plus 21 parameters to ΔΔG data)	Hours	Acc: 72% R: 0.69 Error: N/A	MCC: 0.39 ρ: 0.51 $\Delta\Delta G_{exp,avg}$: -0.12
Linear Interaction Energy (LIE)[b] 2012 [322]	-atomistic -physical potential -fitting (10 parameters to ΔΔG data)	Minutes (estimated)	Acc: 89% R: 0.72 Error: 0.82 (AUE)	MCC: 0.22 ρ: 0.66 $\Delta\Delta G_{exp,avg}$: 0.04
mCSM 2014 [323]	-atomistic -statistical potential -machine learning (72 inputs trained on ΔΔG data)	Seconds	Acc: 82% R: 0.82 Error: 1.0	MCC: 0.39 ρ: 0.53 $\Delta\Delta G_{exp,avg}$: 0.39
DUET 2014 [324]	-meta-prediction using mCSM and SDM	Seconds	Acc: N/A R: 0.71 Error: 1.1	MCC: 0.46 ρ: 0.56 $\Delta\Delta G_{exp,avg}$: 0.26
NeEMO 2014 [325]	-atomistic -statistical potential -machine learning (184 inputs trained on ΔΔG data)	Minutes	Acc: N/A R: 0.77 Error: 1	MCC: 0.45 ρ: 0.62 $\Delta\Delta G_{exp,avg}$: 0.21

ENCoM 2014 [326]	-coarse-grained -physical potential -fitting (4 parameters to ΔΔG data)	Seconds	Acc: N/A R: N/A Error: 1.5	MCC: 0.27 ρ: 0.40 ΔΔG$_{exp,avg}$: -0.50
MAESTRO 2015 [327]	-atomistic and coarse-grained -statistical potential and features -machine learning (9 inputs trained on ΔΔG data)	Seconds	Acc: 82% R: 0.68 Error: 1.1	MCC: 0.50 ρ: 0.67 ΔΔG$_{exp,avg}$: 0.39
SimpleMachine 2015	-atomistic -features -machine learning (4 inputs trained on ΔΔG data)	Seconds	Acc: 82% R: 0.50 Error: 1.5	MCC: 0.23 ρ: 0.48 ΔΔG$_{exp,avg}$: 0.16

Methods which use only sequence information without structural information are not considered here.

Acc, accuracy is reported as a percentage based on the fraction correct. Acc = TP + TN / (TP + FP + TN + FN), where TP, FP, TN, FN represent the number of true positive, false positive, true negative, and false negative cases respectively [328]. Here we consider stabilizing mutations positive and destabilizing mutations negative.

R, Pearson's correlation coefficient [329].

Error, reported in kcal/mol is typically standard error. AUE indicates error is reported as average unsigned error.

MCC, Matthew's correlation coefficient. Calculated as:
MCC = TP*TN - FP*FN / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)] [329].

ρ, Spearman rank order correlation coefficient. This is equivalent to the Pearson's correlation coefficient calculated with ordinal ranking of the data, thus eliminating many problems to do with outliers and anchoring or leverage effects [329].

ΔΔG$_{exp,avg}$, the average effect of all mutations predicted to be stabilizing by the tool, in kcal/mol.

[a] The server for this program is no longer available.

[b] Only semi-automated, but scripts make implementation of these molecular dynamics based procedures somewhat straight-forward.

[c] Estimated based on the needed simulation time of 30 ns with an all-atom explicitly solvated system run in GROMACS on a typical desktop computer.

[d] Estimated based on similarity of the computational procedure to that of FoldX, EGAD, and Eris (without backbone flexibility).

6.4 Results

6.4.1 $\Delta\Delta$G prediction tools do not reliably suggest truly stabilizing mutations

We sought to improve the stability of our previously designed protein, ThreeFoil [44], which possesses resistance to degradation/denaturation by proteases, chemical denaturants, and high temperature, but has only moderate thermodynamic stability (∼6 kcal/mol) [70]. Despite the combined use of numerous stability prediction tools, only 2 of the 10 tested mutations were stabilizing and each by less than 1 kcal/mol.

There exist a considerable number of computational tools which predicting changes in protein stability upon point mutation, and many are automated through a server, standalone program, or readily usable script (Table 6.2). These prediction tools take various approaches to evaluate protein stability. Physical potentials (PP) rely on physical forces such as molecular dynamics forcefields [320, 317, 310, 326, 322]. Statistical potentials (SP) use the probability of observing certain structural arrangements, such as amino acid contact frequencies [316, 313, 307, 311, 319, 314, 327, 324, 325, 318, 323]. Empirical potentials (EP) use a mixture of energy terms from PP and SP with empirical weighting of each term based on experimental data [308, 321]. Some tools rely on particular features, such as change in hydrophobicity or net charge. These features, or heuristics, are used as inputs in machine learning [315, 312, 309], though machine learning has also be applied to SP [327, 324, 325, 318, 323]. The reported performance of these tools is impressive, with overall accuracies (% correctly assigned as stabilizing or destabilizing) >80%, and R-values > 0.7 (Table 6.2).

Using 11 of these tools we tested all 2520 possible point mutations to ThreeFoil (140 structured structurally resolved amino acids * 18 possible amino acids substitutions, not including cysteine). To improve the odds of success we constructed a weighted meta-prediction based on how well each algorithm performed on different kinds of mutations in the Protherm database (see methods and Supplementary Figure 6.5). Meta-predictions have been used in numerous areas of protein science including: covalent modification [330], protein-ligand binding [331], protein-protein binding [332], protein disorder [333, 334], and protein aggregation [335], and generally found to improve the reliability of predictions. We then experimentally validate the top 10 suggested mutations at varying positions (i.e. the top mutations were generally at one or two different positions, but we wanted to diversify the positions we were testing). Determination of $\Delta\Delta$G from kinetic measurements revealed that only 2 of the 10 mutations were unambiguously stabilizing, while 5 were essentially neutral and the remaining 3 were destabilizing (Figure 6.1a, Supplementary Figure 6.6, Supplementary Table 6.4). Notably, the top 3 suggested mutations were destabilizing and the mutation predicted to be the best was, in fact, the worst. These results are much poorer than expected from the reported performance of the tools (Table 6.2), where accuracies of ∼80% give the impression that ∼8 stabilizing mutations would have been obtained. Nevertheless, mutations in Protherm, which are largely fairly conservative, suggest only ∼10% of such mutations would be stabilizing (Figure 6.1b,c,d). Moreover, deep sequencing suggest that as few as ∼1-2% of random mutations are expected to stabilize a protein [336, 337]. Therefore, despite being far from reliable, stability prediction tools can improve the odds of success in protein engineering. Additionally, multi-mutants combining the best mutations were very stabilizing (Supplementary Figure 6.6, Supplementary Table 6.4), though difficult to purify owing to poor solubility, in agreement with predicted aggregation propensity (Supplementary Figure 6.7). Such reduced solubility is likely the result of a bias towards incorporation of hydrophobic mutations on the protein surface.

By surveying the results of numerous experimental attempts to use these same kind of tools to stabilize proteins, we find that success rates are, like ours, typically low (∼25%). The average ΔΔG of mutations predicted to be stabilizing is ∼0 kcal/mol (most are neutral, as in our experiment) (Figure 6.1b), and there is essentially no correlation between predicted and experimentally determined ΔΔG values (Supplementary Figure 6.8). However, the proportion of destabilizing mutations and the average ΔΔG of the mutations are improved over the average of the thousands of mutations found in the Protherm database (Table 6.1, Figure 6.1b,c.d). Moreover, it is extremely rare for a mutation predicted to be stabilizing to be highly destabilizing (< -2 kcal/mol), despite such highly destabilizing mutations being common in the Protherm database (Figure 6.1b), and occurring ∼50% of the time during random mutagenesis [337]. Thus, while stability prediction tools are not reliable when recommending stabilizing mutations, they screen out highly destabilizing mutations and lead to results that are superior to chance.

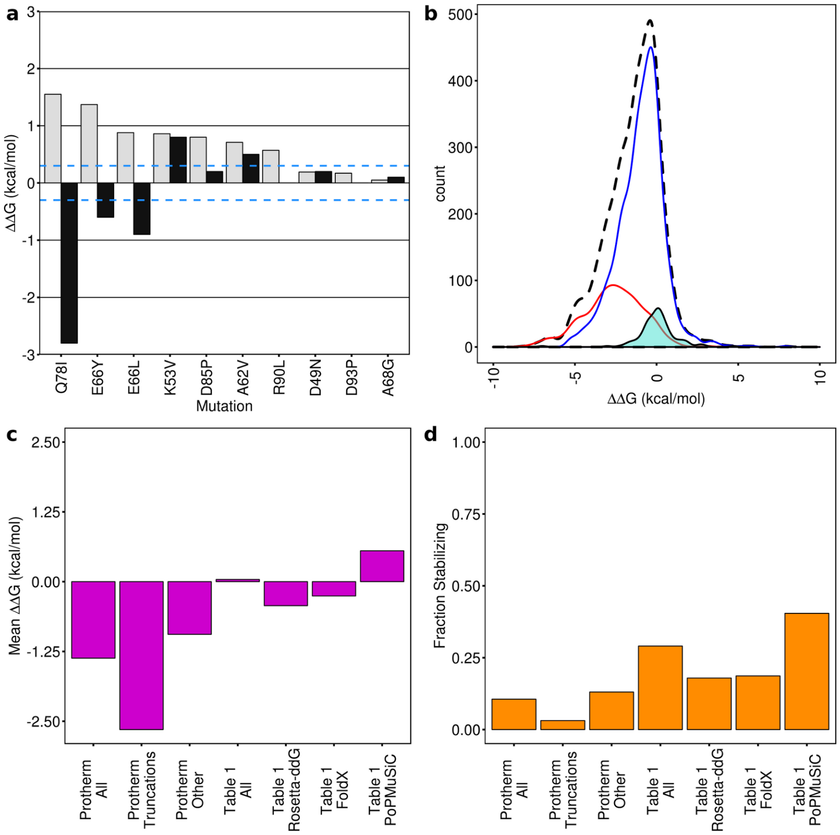


Figure 6.1: Mutations predicted to be stabilizing are most likely to be approximately neutral instead. **(a)** Meta-prediction for mutants of ThreeFoil (grey bars) and experimentally determined changes in stability (black bars), show considerable disparity. Blue dashed lines indicate 0.3 kcal/mol cutoff within which a mutation is considered neutral. **(b)** Density distributions of mutations in the Protherm database (dotted black) as well as sub-populations of hydrophobic truncation mutations (red) and non-truncations (blue). Mutations predicted to be stabilizing by stability prediction tools and subsequently tested, from a number of different groups (Table 6.1), are shown as a filled (cyan) distribution. **(c)** Average impact of mutations for each of the groups shown in **b**, as well as the three tools most commonly employed in experimental engineering attempts (Table 6.1). **(d)** The fraction of mutations which stabilized, shown for the same groups as in **c**.

6.4.2 Algorithms are blind to stabilizing mutations

How can we reconcile the excellent reported performance (Table 6.2) of stability prediction tools with their poor performance in actual engineering attempts (Table 6.1)? Strikingly, we find that known stabilizing mutations are poorly recognized by all tools. As these make up only a small fraction of the datasets used for benchmarking (Figure 6.1b), however, this critical blindness has little impact on overall accuracy. Yet, these mutations are possibly of the greatest interest in biotechnology.

By testing 21 automated tools (Table 6.2) against a curated dataset of ∼600 mutations not used in the tool's training or parameterization (see Methods, Supplementary Figure 6.9) we confirm that while overall accuracies are indeed ∼80%, stabilizing mutations are only correctly recognized ∼50% of the time, comparable to a coin-toss (Figure 6.2a, green bars). Furthermore, of mutations predicted to be stabilizing, the fraction that truly are (positive predictive value, PPV or Success rate) is about 0.5 (Figure 6.2b, Table 6.2). Tellingly, for many tools, the average experimental change in stability for mutations predicted to be stabilizing ($\Delta\Delta G_{avg,exp}$), is actually destabilizing (Figure 6.2c) with the majority of recommended mutations being neutral (Supplementary Figure 6.10). Thus, prediction tools are not only blind to stabilizing mutations — being unable to recognize them when presented — but those they do recommend are most likely neutral.

Are stabilizing mutations just intrinsically more challenging to predict than their destabilizing counterparts? Because destabilizing mutations tend to have a greater absolute impact on stability than stabilizing mutations (Figure 6.1b), stabilizing mutations might be challenging to identify owing to their subtlety. In fact, moderately destabilizing mutations (-2 to 0 kcal/mol) are still more accurately predicted than moderately stabilizing ones (0 to +2 kcal/mol), despite having effects of the same magnitude (Figure 6.2a, yellow vs. green bars). Examining the bias between recognition of moderately stabilizing and destabilizing mutations (Figure 6.2a, difference between yellow and green bars) highlights a few tools with a low bias: DFire2, SDM, and MultiMutate. These tools were developed without training or extensive parameterization using Protherm (Table 6.2), often relying on known $\Delta\Delta G$ data simply to generate a conversion factor from internal energy units to kcal/mol. Other tools which similarly limit Protherm usage (TI, Eris, Rosetta) — being largely developed from structural information — are less biased than average. Importantly, the aforementioned tools, with the exception of MultiMutate, have average if not above average, overall performance (Figure 6.2c,d), and may represent the kind of tools that have the highest potential for future improvement. By contrast, tools which rely heavily on Protherm data for development are strongly biased against stabilizing mutations, with machine learning tools showing particularly high inclination to predict stabilizing mutations as destabilizing, a known problem of training against data that is biased towards a particular class (destabilizing mutations in the case of Protherm) [338].

In developing IMutant3, Capriotti *et al.* attempted to correct the bias against stabilizing mutations in Protherm by assuming that for each mutation X to Y, there must be a mutation Y to X, of equal magnitude but opposite sign [315]. This approach, however, has yielded only a small improvement over the earlier IMutant2 and fairly poor performance overall (Figure 6.2). Most stabilizing mutations are not simply the opposite of destabilizing mutations. For instance, almost all highly destabilizing mutations in the Protherm database are truncations of a larger hydrophobic amino acid to a much smaller one, leaving a large void (Figure 6.1b,c). The reverse, where the protein is stabilized by filling a large void, is rarely seen in natural or well-designed proteins, as they do not have such obvious void spaces in the first place (though filling of smaller packing deficiencies has been successful [339]).

Thus, the inability of prediction tools to identify truly stabilizing mutations — a crucial problem for protein engineering where these are precisely the mutations being sought — may be minimized by developing tools using structural data, which has no particular bias towards destabilizing mutations.

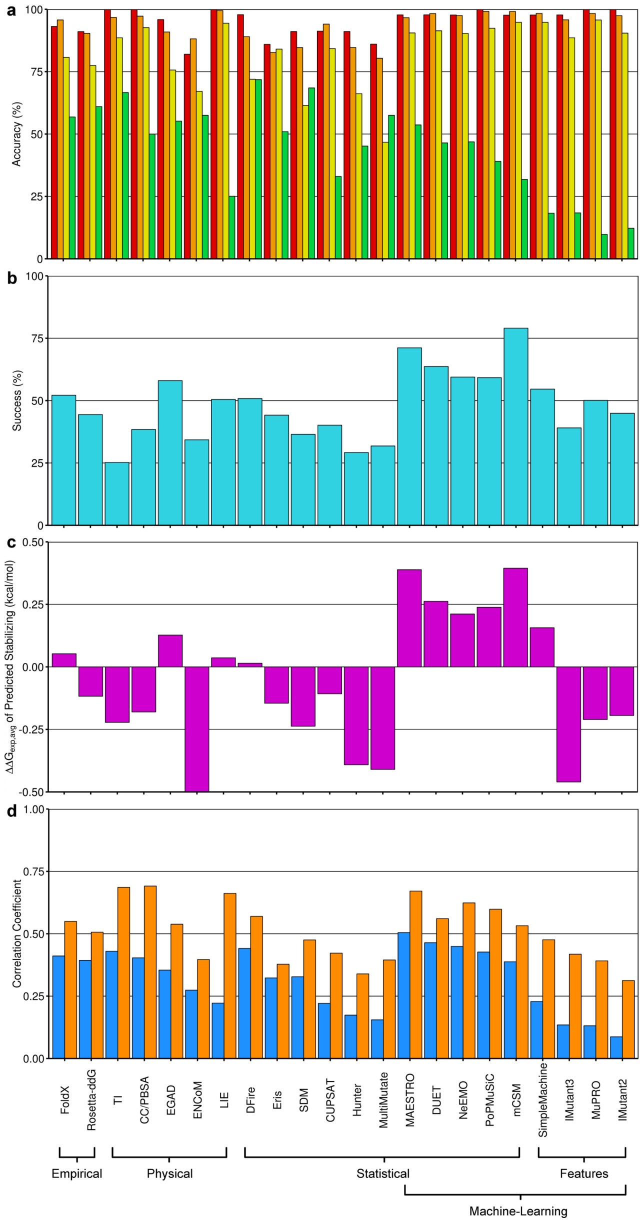


Figure 6.2: Analysis of stability prediction tools. **(a)** Accuracy (percent correctly classified as stabilizing/destabilizing) binned by experimentally determined stability: Extremely destabilizing (red bars, -6 to -4 kcal/mol), highly destabilizing (orange bars, -4 to -2 kcal/mol), moderately destabilizing mutations (yellow bars, -2 to 0 kcal/mol), and moderately stabilizing mutations (green, 0 to +2 kcal/mol). **(b)** Success rate or positive predictive value (PPV) shows how often a mutation predicted to be stabilizing by a tool, actually is experimentally. **(c)** The average ΔΔG of all mutations predicted to be stabilizing by each tool. **(d)** Matthews (blue) and Spearman (orange) correlation coefficients, which measure correlation of binary classification and rank order respectively, shown for each tool.

6.4.3 Commonly reported performance metrics may be misleading

The most commonly reported performance metrics for stability prediction tools are: accuracy, R, and standard or average unsigned error (Table 6.2). Unfortunately, the bias towards destabilizing mutations in the Protherm database — consistently used for testing performance — means that accuracy is an inappropriate measure, since high accuracy ($\sim$85%) can be achieved simply by predicting all mutations as destabilizing. While an error of 0 kcal/mol would represent a perfect prediction, mutations in the Protherm database predominantly occur in the range of -2.5 to 0.5 kcal/mol and as such consistently predicting all mutations as being destabilizing by $\sim$1 kcal/mol can yield competitive errors (see below). Therefore while error is not necessarily an inappropriate metric in this case, it poorly discriminates between methods given the current quality of the prediction tools. The proper calculation of R should involve removal of outliers which strongly affect this value [329], but there are seldom good reasons to suspect certain mutations as being outliers to the remaining population (e.g. a poorly resolved crystal structure, or mutations involving a particular amino acid), and consequently many reports identify "outliers" as the points which are most poorly predicted, making the interpretation of this measure difficult. Using mock predictions as well as an intentionally over-simplified machine learning tool, we show that the Matthews and Spearman correlation coefficients (MCC and ρ) are robust and appropriate measures of performance when testing stability prediction tools and should be adopted over accuracy, R, and error.

From the Protherm database, we collected $\sim$1000 point mutations as a training set that is not part of the $\sim$600 mutation testing set. We considered a simple "predictor" that assigns the average $\Delta\Delta$G from the training set to any test mutation it is given. This indiscriminate strategy nevertheless produces an accuracy and standard error that are comparable to the published prediction tools, while all correlation coefficients reveal the expected poor performance (Supplementary Figure 6.11a). If instead, when asked to predict glycine and alanine scanning mutations the average $\Delta\Delta$G in the training set for these types of mutations is used, an R comparable to some of the poorer published tools is obtained, while MCC and ρ continue to reveal the limited predictive capability (Supplementary Figure 6.11b).

In order to provide an estimate of the kind of performance expected from a "barebones" prediction, we developed our own intentionally mediocre prediction tool based using machine learning with only 4 broad features (change in amino acid size, change in amino acid polarity, secondary structure propensity, and solvent accessible surface area, see methods). This "SimpleMachine" nevertheless achieves a very high accuracy (83%), low error

(1.5 kcal/mol) and average R (0.50) (Supplementary Figure 6.11c). Shockingly, based on these most commonly reported metrics, it outperforms more than half of the published tools (Table 6.2, Supplementary Figure 6.9). The MCC metric, however, reveals that SimpleMachine is one of the poorer performing tools, as expected (Figure 6.2e). While ρ for SimpleMachine is approximately average it is in the bottom half of the tools. Thus, while accuracy and measures of error can be unreliable performance metrics for stability prediction, MCC and ρ are robust. Furthermore, both MCC and ρ are little impacted by outliers — unlike R — simplifying their use and making them less variable in the hands of different users [329]. A comparison of the tools based on the date of publication demonstrates that progress has been slow and uncertain, likely hobbled by the use of uninformative metrics (Figure 6.3). As such, the adoption of MCC and ρ as the standard performance metrics should allow the most promising tools to be distinguished, highlighting the best paths towards future improvement.

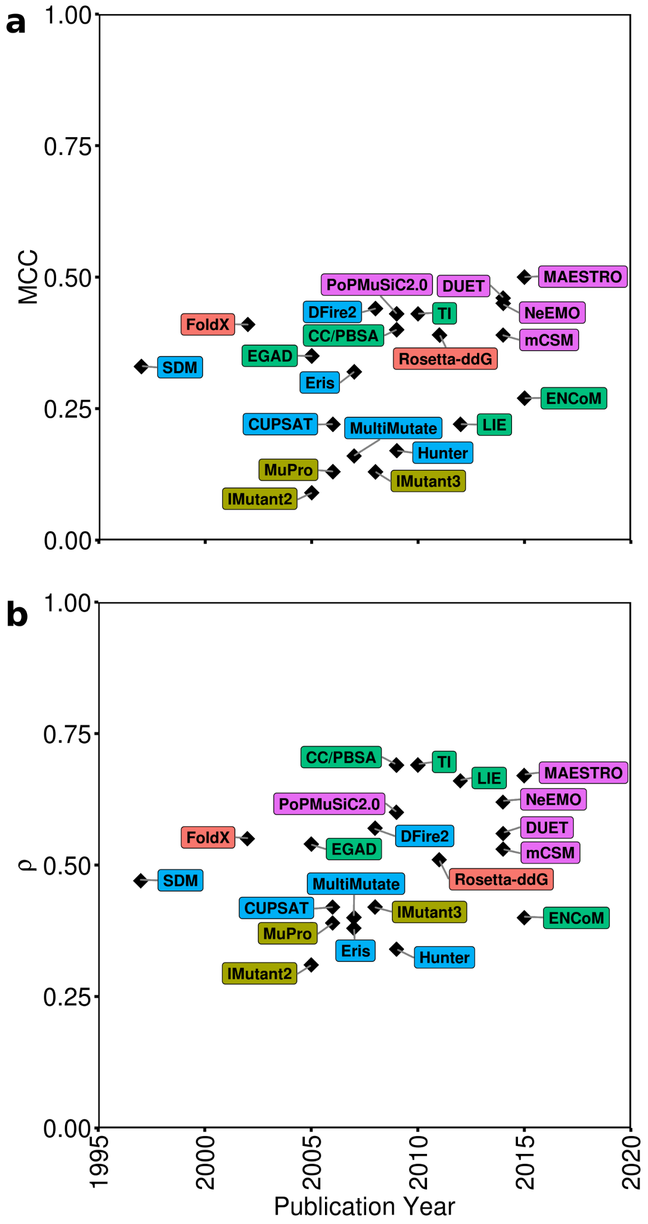


Figure 6.3: Performance of stability prediction tools over time. Matthews (MCC) **(a)** and Spearman (ρ) **(b)** correlation coefficients are shown for each algorithm as a function of publication date. While there is some evidence for improvement over time (particularly as measured by ρ), this improvement is highly variable. The basis of the tools (physical-green, empirical-red, statistical without machine learning-blue, statistical with machine learning-magenta, and heuristic or feature-based with machine learning-yellow) are indicated.

6.4.4 Recommended strategy for choosing stabilizing point mutations

While our analysis does not promote confidence in stability prediction tools when used to enhance protein stability, an optimum strategy yielding useful results may still be possible. Previous use of stability prediction tools to engineer stabilizing mutations (Table 6.1, Supplementary Table 6.3) shows that for tools with experimental validation, success rate is highest for PoPMuSiC, followed by FoldX and Rosetta-ddG. Based on the metrics we have identified to be the most useful: $\Delta\Delta G_{avg,exp}$, MCC and ρ (Figure 6.2c,d), the same relative ranking is seen. This suggests that performance against the Protherm database is a useful proxy to performance when engineering enhanced stability. Thus, if using a fully automated tool, the best choices come from those based on statistical potentials and machine learning: MAESTRO, DUET, NeEMO, and PoPMuSiC.

The use of simple equilibrium molecular dynamics to identify problems with suggested mutations (Table 6.1, +MD) yields a small but tangible benefit when used as a second pass after initial selection by FoldX and/or Rosetta-ddG, equivalent to a boost of ~0.5 kcal/mol in $\Delta\Delta G_{avg,exp}$. However, the use of MD with alchemical transformations — for instance the TI method partially automated by Seeliger *et al.* [320, 340] — when applied as a second pass after initial selection by FoldX [341] gave a considerable (~1 kcal/mol) boost, and the highest success rates seen experimentally (Table 6.1, FoldX+TI).

Therefore, both previous experimental work and our analysis of tool performance suggest an optimum workflow where: MAESTRO, DUET, NeEMO, PoPMuSiC are used as an initial screen (either alone or together with a simple majority vote), followed by the more computationally expensive MD or TI to maximize the probability of success. It may also be recommended to remove mutations likely to reduce solubility (even if they would improve stability) such as mutations to more hydrophobic residues on the surface (Figure 6.4).

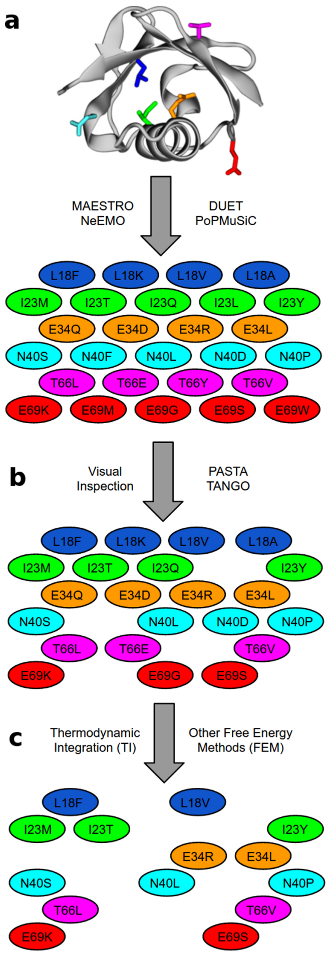


Figure 6.4: Suggested workflow to stabilize a protein of interest. Based on our analysis (Figure 6.2) and the results of other experimental groups (Table 6.1, Supplementary Table 6.3), a workflow for generating a small subset of mutations likely to stabilize a protein of interest is shown. **(a)** The wild-type structure is given to the most successful stability prediction tools (MAESTRO, NeEMO, DUET, PoPMuSiC), to predict the effect of all (19 x N) point mutations to each residue, producing a set of potentially stabilizing point mutations (shown as colored ovals with hypothetical mutations written inside). Note that while a single tool can be used, owing to the high correlation between these tools (Supplementary Figure 6.12), majority voting could improve performance [342]. **(b)** Either by visual inspection (to remove mutations which produce hydrophobic surface patches) or through the use of automated tools for predicting aggregation propensity (Supplementary Figure 6.7), mutations which might compromise solubility are eliminated. **(c)** Finally, molecular dynamics including free energy methods such as thermodynamic integration (TI) — which are computationally expensive — are used to look for structural abnormalities/weaknesses introduced by the mutations, or in the case of TI, to provide a successful and orthogonal prediction method. This step is performed last such that the high computational cost is offset by testing the fewest possible number of mutations.

6.5 Discussion

Stabilizing mutations confer a host of valuable attributes to a protein: improving capacity for evolution or design of novel functions, greater chance of improving existing function [304, 153], and improved thermal stability [343]. Higher thermal stability allows for elevated working temperatures and efficiencies in biocatalysts [299, 34] and longer shelf-life and reduced immunogenicity of biotherapeutics [301, 302, 303]. Despite the well-recognized value of predicting changes in protein stability upon mutation, and the substantial body of work devoted to developing tools for these predictions (Table 6.2), the capability to accurately predict stabilizing mutations is limited as seen in our engineering attempts presented here (Figure 6.1a) and those of other groups (Table 6.1, Supplementary Table 6.3). This limits the engineering of proteins with ideal qualities for biotechnological applications.

The Protherm database, while a valuable resource for studying the effect of mutations on stability, is heavily biased towards destabilizing mutations, in particular truncation mutations. This bias reflects the reality that most mutations to natural or well-designed proteins are destabilizing [34, 33]. Nevertheless, the bias impacts not only the development of stability prediction tools, but how their performance should be tested. Critically, the simple metric of accuracy (fraction correct) is not meaningful in this case, yet is the most frequently reported, giving the false impression of excellent performance. That the majority of stability prediction tools (and particularly those trained or parameterized using Protherm) are better at predicting destabilizing mutations highlights how use of this inappropriate metric may have slowed development (Figure 6.3). Because experimental methods are not perfect and stability prediction is still developing, all tools have standard errors which can be mimicked simply by making highly conservative estimates. Even R-values may be misleading when a few extreme cases are somewhat well predicted or when the most poorly predicted cases are removed as “outliers”. Conversely, MCC and ρ offer robust metrics of performance success, and more specific metrics like success rate (PPV) and the average experimental ΔΔG of mutations predicted to be stabilizing ($\Delta\Delta\mathrm{G}_{avg,exp}$), are particularly useful for protein engineers seeking to understand the likely outcomes of applying these stability prediction tools to their proteins.

Viewing stability prediction tools through the lens of the MCC and ρ reveals that methods based primarily on heuristics or features (MuPro, IMutant2, IMutant3 and SimpleMachine) perform very poorly, despite their high accuracy, potentially low standard error, and average R-values. By contrast, tools based on statistical potentials that employ machine learning (MAESTRO, DUET, NeEMO, PoPMuSiC, and mCSM) perform much better, yet are still in need of improvement. It may be critical for future development to note that tools trained on structural, rather than stability data (DFire, Eris, and to a large extent Rosetta-ddG [321]), show competitive performance, while minimizing the bias against recognition of stabilizing mutations, and thus hold considerable promise. A similar case can be made for tools using molecular dynamics forcefields, which are often parameterized on small molecules in the gas phase (TI, CC/PBSA, EGAD, LIE) and thus have no *a priori* reliance on stability data. In fact, molecular dynamics may be used alone to suggest stabilizing mutations [344] or as a method for generating conformational variation which may improve performance of existing stability prediction tools [345, 346].

Therefore, our detailed analysis of stability prediction tools not only highlights weaknesses in these tools (Figure 6.2), but provides useful metrics for benchmarking success and future development (Figure 6.3). Additionally we use this analysis to suggest a workflow for protein engineers seeking a current best practice for improving protein stability in an automated manner that minimizes human and wet-lab resources (Figure 6.4).

6.6 Methods

6.6.1 Protherm dataset construction

Single point mutations to proteins of known structure were taken from the Protherm database [306] (http://www.abren.net/protherm/). The criteria for inclusion in our dataset follow. The ΔΔG must have been extrapolated to, or determined in, a solution with 0 M denaturant. The experiment must have been performed in the temperature range (20 to 30 °C), or could be reliably extrapolated to within this range. The experiment must have been performed at a pH between 5.0 and 9.0. The protein could not contain prosthetic groups such as heme (as these cannot be taken into account by most tools). Similarly the experiment could not have been conducted in the presence of ligands. For mutations determined by more than one method, the average value was taken.

Because entry of data with the incorrect sign is a known problem with the Protherm database, entries were scanned manually and automatically. Automatic scanning involved searching for numerous entries from the same publication that were all stabilizing (since this is rare). Corrections to the data were only performed after verifying with the original publication.

Mutations used for training various tools were removed in order to provide a fair test for all methods, leaving 605 single point mutations in total. In the case of: CC/PBSA, LIE, and TI, the published results were used rather than retesting on our 605 mutation dataset because CC/PBSA recently became unavailable for use, and LIE and TI, while having some scripts are not entirely automated at this point. Since these methods all provide a glimpse of what is possible using physical potentials, we chose to include them in order to better evaluate which approaches will be best for future development.

6.6.2 Meta ΔΔG prediction

Using our 605 mutation dataset, we ran predictions for each point mutation through 11 tools: CUPSAT, DFire, EGAD, FoldX, Hunter, IMutant3, MultiMutate, MuPRO, PoPMuSiC, Rosetta-ddG, and SDM. The point mutations were categorized based on five classes: change in polarity (less polar, the same polarity, more polar), change in size (smaller, the same size, larger), solvent accessible surface area (SASA) of the WT residue (buried, partially exposed, fully exposed), secondary structure at the site of mutation (alpha, beta, turn, unstructured), and if the WT residue was a glycine or not. Polarity measures were taken from the work of Wimley *et al.* (see Table 2, 3rd column within the citation) [347] and a difference of at least 0.1 kcal/mol (solvation free energy) was needed otherwise the mutation was considered the same polarity. Size measurements were taken from the work of Darby and Creighton [348] and a difference of at least 19 Å^3 (the difference between glycine and alanine, or approximately a CH_3 group) was used otherwise the mutation was considered the same size. SASA for a given residue was calculated by VMD (http://www.ks.uiuc.edu/Research/vmd/) based on the WT PDB structure, and a relative SASA was computed by dividing by the SASA of that residue alone. Residues were considered buried when the relative SASA was ≤ 0.05, partially exposed when > 0.05 and ≤ 0.2 and fully exposed when > 0.2. Note the ratio of 0.2 as the cutoff for exposed may appear small, but the backbone atoms are considered here also, such that a fully exposed sidechain within the context of a beta-sheet could have a fairly low ratio. The secondary structure class of the residue was calculated using DSSP [349] based on the WT PDB structure.

The performance of each prediction tool on each type within each class was calculated as the MCC when considering only mutations in that type (Supplementary Figure 6.5a-e). In the event of a negative correlation coefficient, a value of 0 was used instead. This was done because we assumed that negative correlations (which were at most -0.1) were simply the result of noise and not a true anti-prediction, and thus the prediction the given tool on that particular class of mutations was simply ignored. The meta prediction for putative mutations to ThreeFoil was computed by simply averaging the predictions of the 11 tools, each weighted based on the average MCC values obtained for that class and type of mutation on the 605 mutation dataset.

For determining the expected performance of the meta-prediction and how it increased as more tools were added (Supplementary Figure 6.5f), the meta-prediction MCC-weights were calculated on ∼half of the 605 mutations and then tested on the remaining half. The procedure was repeated 1000 times with halves being selected randomly and the average values reported. For different numbers of tools used, the correlation coefficients (Supplementary Figure 6.5f) are computed as the average of all possible tool combinations.

6.6.3 Expression and purification of ThreeFoil

The transformation, expression, and purification of WT ThreeFoil and all mutants was performed exactly as reported in Chapters 2 and 5 [44, 70].

6.6.4 Folding and unfolding kinetics

Kinetic measurements for WT ThreeFoil and all mutants were performed exactly as reported in Chapter 5 [70], except that a Tris buffered solution (150 mM NaCl, 50 mM Tris, pH 8.1) was used instead of phosphate buffered solution.

6.6.5 Statistical analysis

All statistics were calculated by taking 70% of the dataset at random, computing the statistic of interest, and repeating the procedure 1000 times in order to obtain a sound estimate of the statistics. The average value is reported. Binary statics are calculated as in [328], and correlation coefficients are calculated as in [329].

For statistics using binary classification (Accuracy, MCC, positive predictive value or Success Rate) only mutations with an experimentally determined $\Delta\Delta$G > 0.3 or $<$ -0.3 kcal/mol were used. Similarly, for $\Delta\Delta G_{exp,avg}$ only mutations predicted to stabilize by > 0.3 kcal/mol were considered when calculating the average experimentally determined $\Delta\Delta$G of those mutations. Along similar lines, when determining the success or failure of prediction algorithms (Table 6.1 and Supplementary Table 6.3) only mutations predicted to be stabilizing by > 0.3 kcal/mol were included when calculating the final aggregate success metrics (unless a mutation predicted to be stabilizing by > 0 but < 0.3 kcal/mol was successful, in which case that was included). The cutoff value was suggested by Pokala *et al.* in their development of EGAD, [310] by comparing $\Delta\Delta$G values for the same mutation obtained through different experimental techniques [350, 351]. This cutoff was used primarily to avoid penalizing the prediction tools in cases where it was in fact the experiment that was incorrect. For instance, if a mutation was predicted to stabilize by 0.5 kcal/mol, but experiment showed destabilization at -0.1 kcal/mol, it could easily be the case that the true experimental value is as high as 0.2 kcal/mol and thus counting this an a failed prediction will lead to needless noise and unreasonably poor performance metrics. Since the primary objective was to show that these prediction tools do not perform well, giving them every benefit of the doubt ensures that such a conclusion is justified. Similarly, if a tool predicted a mutation to stabilize by only 0.1 kcal/mol and yet the experimental validation showed a clear destabilization (e.g. -1.0 kcal/mol) this was also not counted as a failure, since it would be unreasonable to make mutations predicted to be so marginally stable unless no other options were present.

The original meta-prediction used for choosing mutations to ThreeFoil was performed separately from the subsequent analysis of Protherm and the prediction tools. Therefore the meta-prediction uses only 11 tools rather than the full 21 tested here, and does not make use of the aforementioned cutoff, instead counting any mutation > 0 kcal/mol as stabilizing and ≤ 0 kcal/mol as destabilizing.

6.6.6 SimpleMachine construction and training

Simple Machine was constructed using random forest regression as implemented in the scikit-learn python package (http://scikit-learn.org/stable/). Random forest regression parameters such as: number of estimators, tree depth, and others (see the "RandomForestRegressor" package of scikit-learn) were optimized through 200 trials using the hyperopt python package (https://github.com/hyperopt/hyperopt).

The inputs for each mutation were: change in amino acid size, change in amino acid polarity, SASA, and secondary structure propensity. The values were these inputs were determined in the same way as the meta $\Delta\Delta$G prediction. The value for regression was the $\Delta\Delta$G of the mutation. 1000 single point mutations were used for training that were not part of the 605 mutations used for testing. Both sets of mutations come from all structural classes and should be fairly similar in distribution.

6.7 Supplemental Information

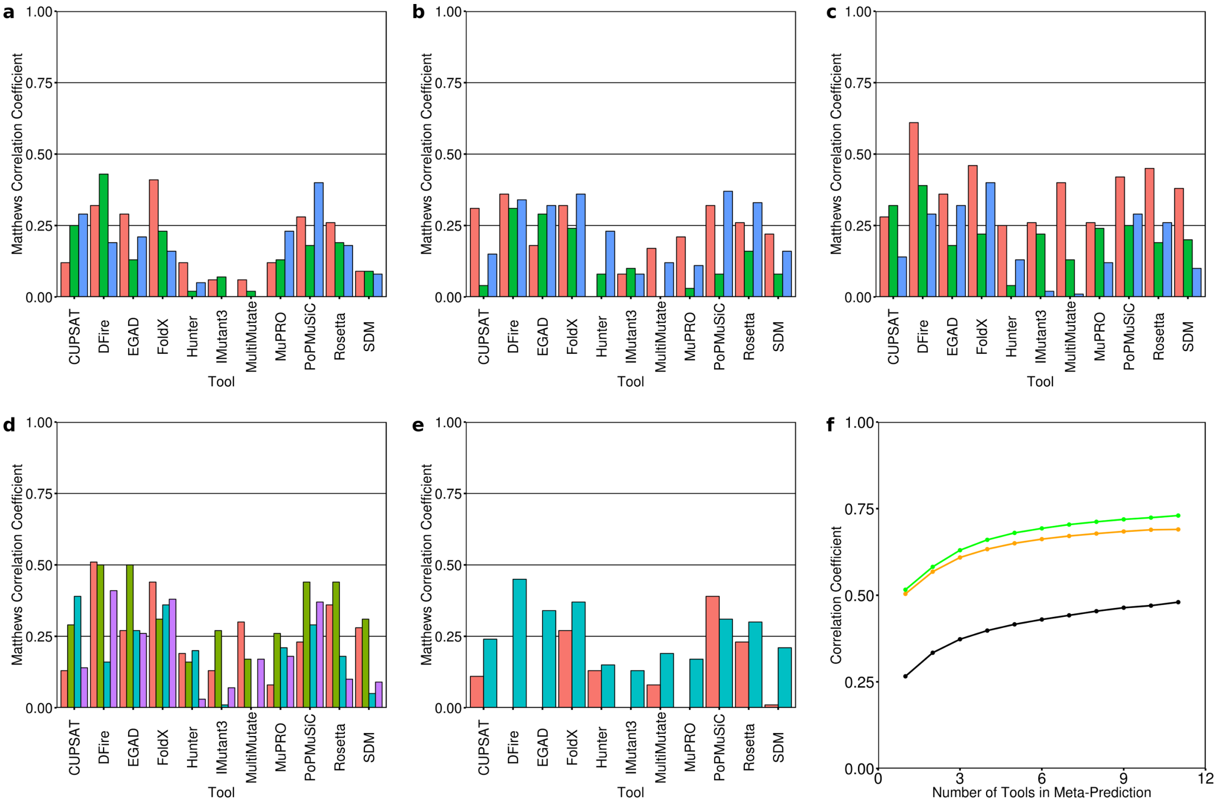


Figure 6.5: Meta-predictions of ΔΔG. Performance of each tool used is shown on different classes of point mutations as measured by MCC. **(a)** Polarity, with types: less polar (red), approximately the same polarity (green), and more polar (blue). **(b)** Size, with types: smaller (red), approximately the same size (green), and larger (blue). **(c)** Solvent accessible surface area of the WT-residue, with types: buried (red), partially exposed (green), and exposed (blue). **(d)** Secondary structure of the backbone around the WT-residue, with types: helical (red), beta (green), turn (cyan), unstructured (purple). **(e)** Whether the mutation is to or from a glycine (red) or non-glycine (cyan). **(f)** The performance of each tool on a training set of mutations was used to produce a meta-prediction as a weighted average using the correlation coefficients in (a,b,c,d,e). As more of the tools were incorporated, there was a continuous improvement in meta-prediction performance measured not only by MCC (black) but ρ (orange) and *R* (green) correlation coefficients also.

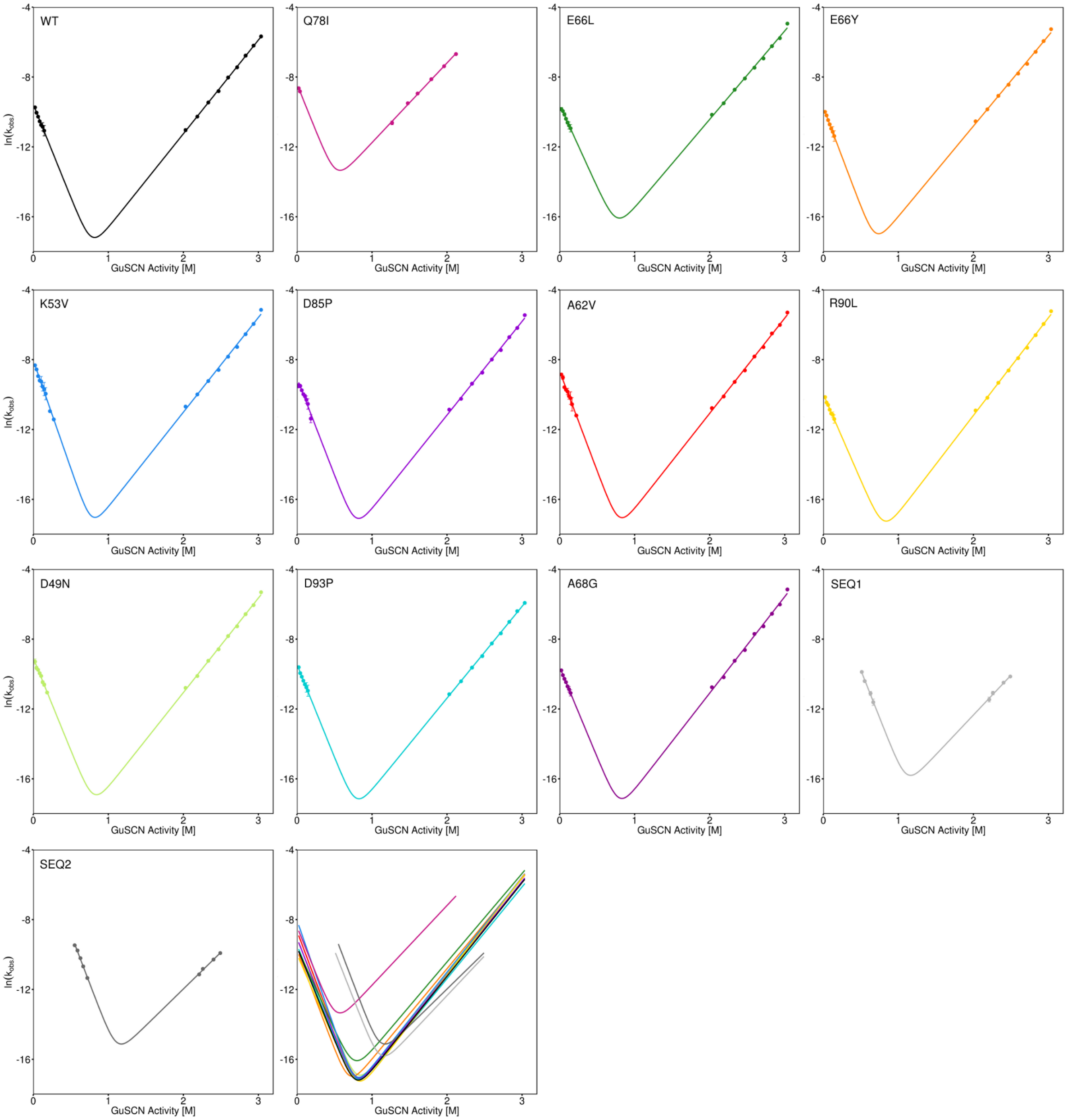
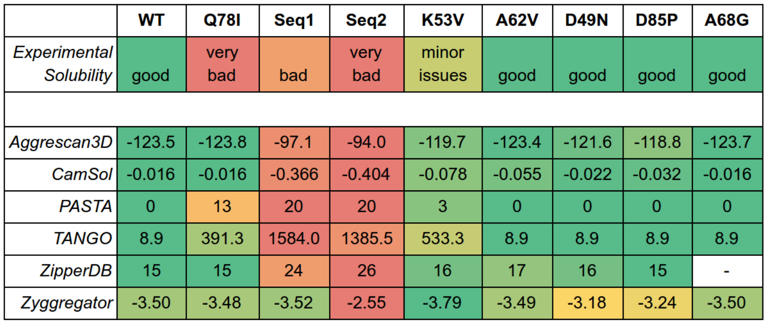


Figure 6.6: Chevron plots of ThreeFoil mutant kinetics. Kinetic measurements are shown for WT and each point mutant, as well as the multi-mutants (see Supplementary Table 6.4 for details of the fits). The final plot at the bottom right shows all mutants together as simply the chevron fit (without experimental data-points) using the same color as in the individual plots.



	WT	Q78I	Seq1	Seq2	K53V	A62V	D49N	D85P	A68G
Experimental Solubility	good	very bad	bad	very bad	minor issues	good	good	good	good
Aggrescan3D	-123.5	-123.8	-97.1	-94.0	-119.7	-123.4	-121.6	-118.8	-123.7
CamSol	-0.016	-0.016	-0.366	-0.404	-0.078	-0.055	-0.022	-0.032	-0.016
PASTA	0	13	20	20	3	0	0	0	0
TANGO	8.9	391.3	1584.0	1385.5	533.3	8.9	8.9	8.9	8.9
ZipperDB	15	15	24	26	16	17	16	15	-
Zyggregator	-3.50	-3.48	-3.52	-2.55	-3.79	-3.49	-3.18	-3.24	-3.50

Figure 6.7: Aggregation propensity prediction of ThreeFoil mutants. Qualitative observations of solubility during purification and kinetics experiments ("Experimental Solubility") are compared with predictions from several automated aggregation prediction tools: Aggrescan3D [352], CamSol [353], PASTA [354], TANGO [355], ZipperDB [356], Zyggregator [357]. Many tools correctly identify that the multi-mutants (Seq1 and Seq2) are less soluble than the single mutants. Only PASTA and TANGO appear to predict some of the very poor refolding behaviour of Q78I and minor refolding problems of K53V.

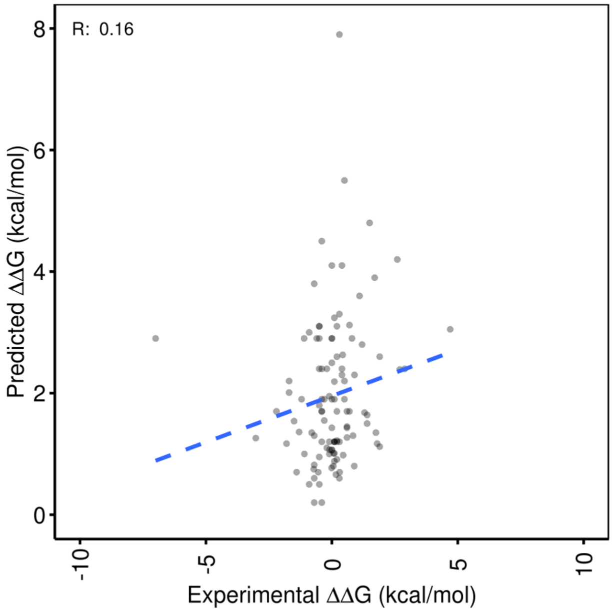


Figure 6.8: Predicted stability does not correlate with experimental when designing stabilizing mutations. The predicted stability of mutations chosen with the goal of stabilizing a given protein (Table 6.1, Supplementary Table 6.3) are compared with their subsequent experimentally determined changes in stability. Despite predictions across a range of stability increases, the actual impact is a fairly narrow distribution centered around neutrality and no significant correlation.

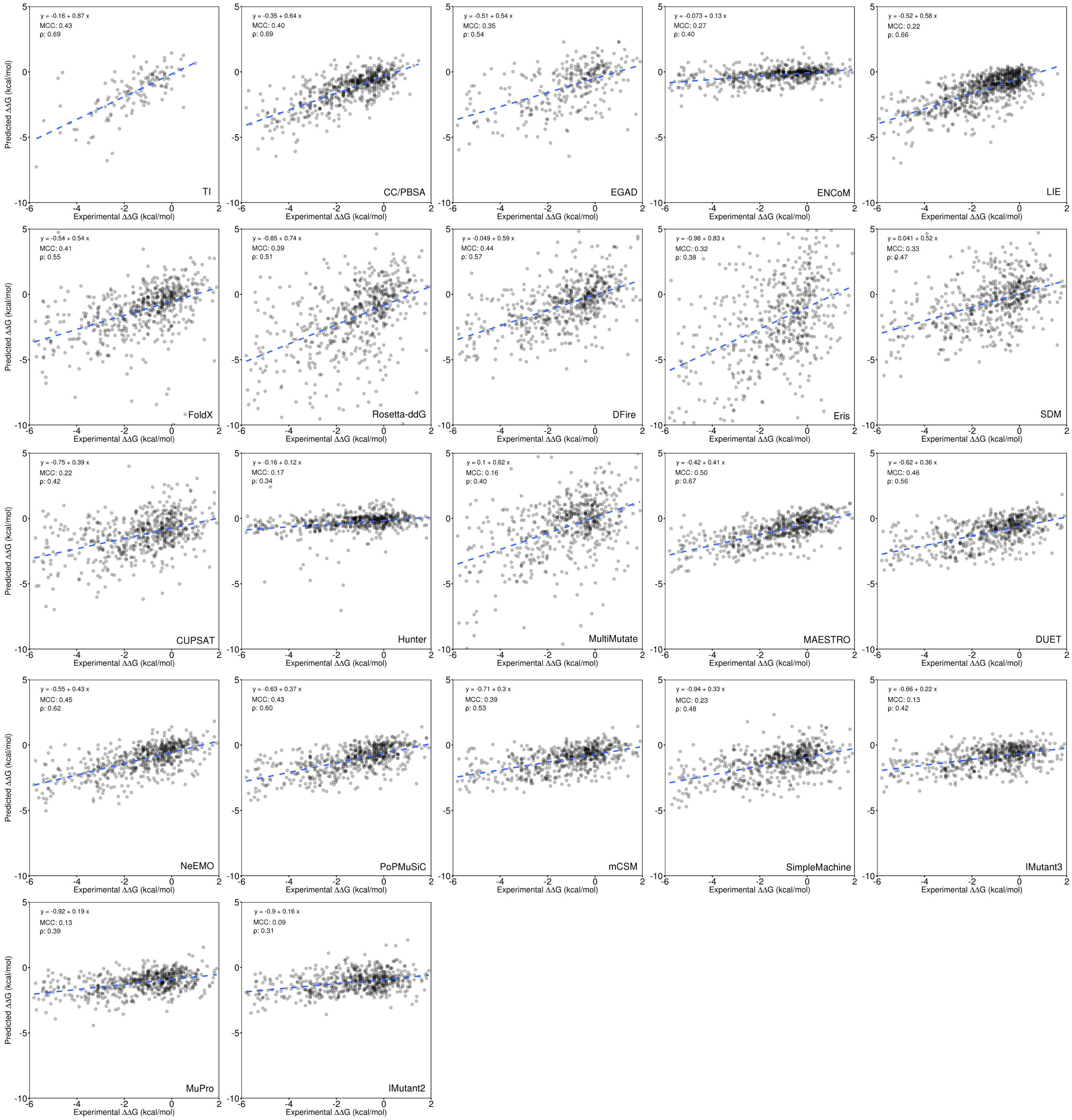


Figure 6.9: Individual predictions on the known mutation dataset for all tools. Scatter plots of the predicted *versus* experimentally determined change in stability are shown for each tool as well as the line of best fit and its equation, and the Matthews (MCC) and Spearman (ρ) correlation coefficients.

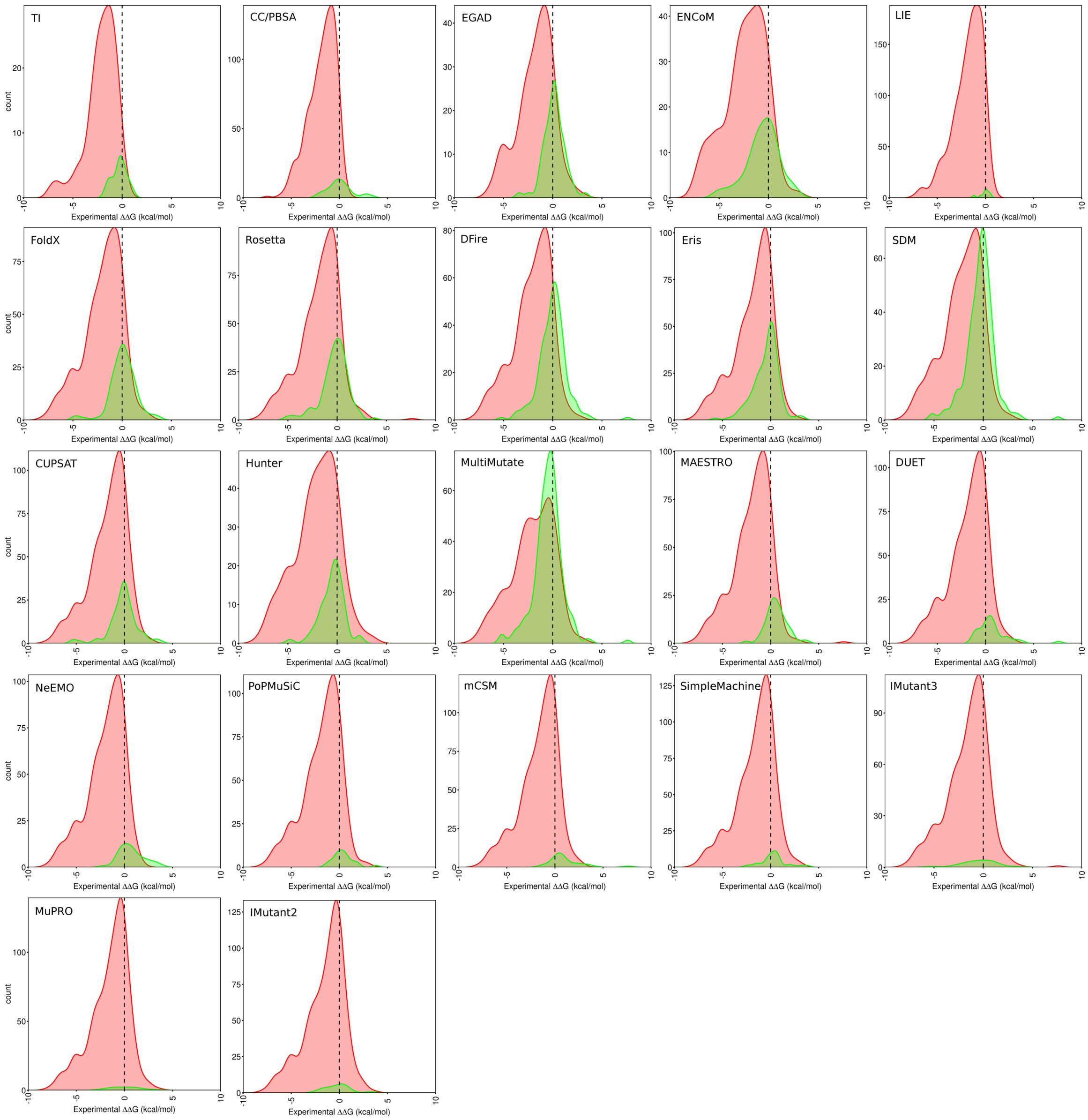


Figure 6.10: Distribution of mutations categorized by each tool. The experimental ΔΔG of mutations in the ∼600 mutation test dataset is shown based on each tool's prediction as destabilizing (red) or stabilizing (green). Algorithms we identified as the best (e.g. MAESTRO, DUET, NeEMO, PoPMuSiC) (Figure 6.3) show a green distribution centered around slightly stabilizing values experimentally. While methods such as FoldX and Rosetta-ddG do not perform as well as MAESTRO and have their predicted stabilizing (green) distributions centered directly at 0 kcal/mol, they predict a much larger number of stabilizing mutations, giving more options for engineering. In stark contrast, feature and/or heuristics based tools (SimpleMachine, IMutant2, IMutant3, MuPRO), which we indicated perform poorly (Figure 6.3), not only have green distributions centered around 0 kcal/mol, but recommend very few mutations.

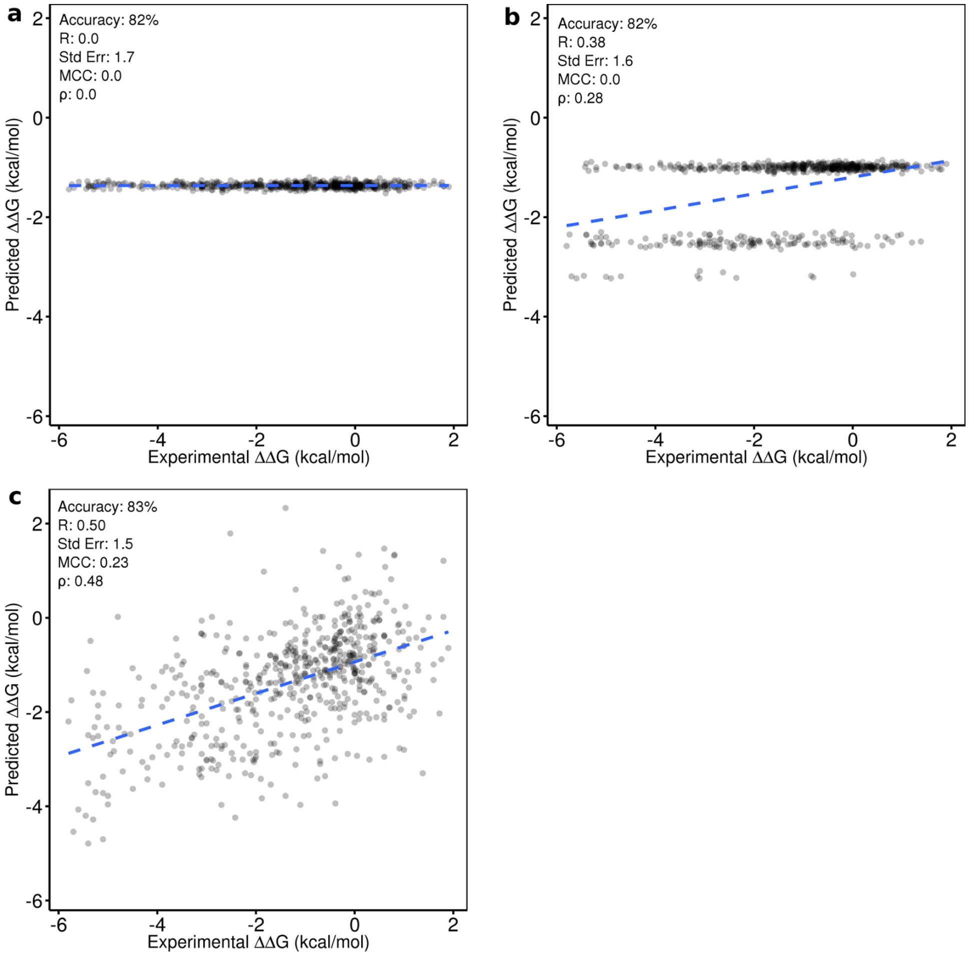


Figure 6.11: Mock predictors and SimpleMachine. **(a)** All mutations in the test set (605 mutations) are predicted to have a ΔΔG of -1.36 kcal/mol, the average of mutations in the training set (1063 mutations). This demonstrates that an accuracy and standard error comparable to published tools can be achieved using the most crude of predictions, revealing the limited value of these performance metrics for this kind data. **(b)** Alanine or Glycine scanning mutations in the test set are predicted to have a ΔΔG of -2.52 kcal/mol and -3.18 kcal/mol, respectively, the average of those kinds of mutations in the training set. Mutations which are neither Alanine nor Glycine scanning are predicted to have a ΔΔG of -1.00 kcal/mol, the average value of those mutations in the training set. This “mock predictor” demonstrates that r can be dramatically boosted, into the range of published tools, by very simple classification of mutation types. (c) An intentionally simple and poorly predictive machine learning tool, called “SimpleMachine”, which relies on only 4 classifications as input features is able to achieve highly competitive accuracy, *R*, and standard error, but a poor MCC and modest ρ reveal its true lack of predictive power.

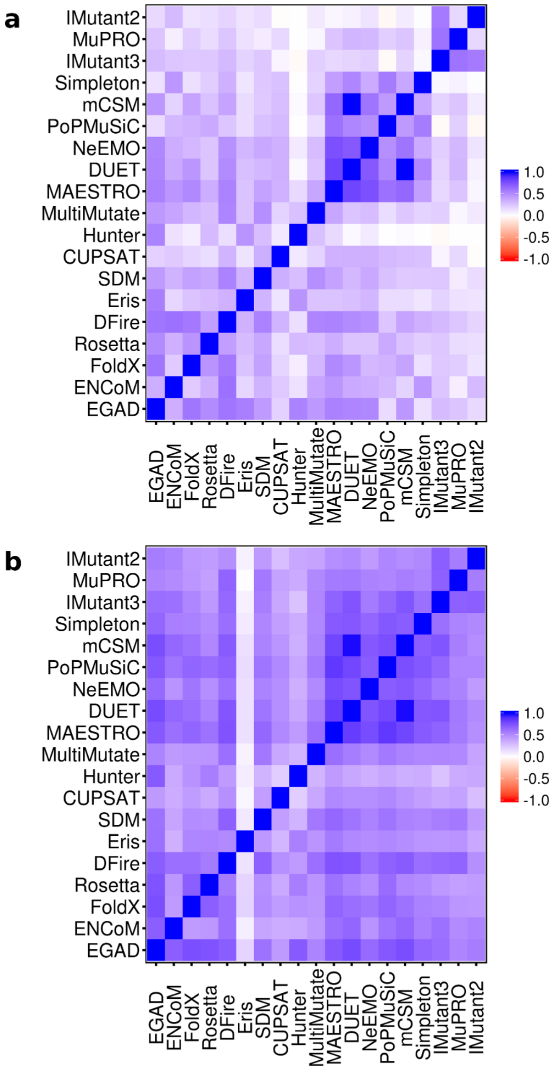


Figure 6.12: Cross-correlation matrix of the stability prediction tools. Note that CC/PBSA, LIE, and TI are not shown here as those were tested on different datasets (see original reports for those methods, Table 6.2). The Matthews **(a)** and Spearman **(b)** correlation coefficients are shown and color-scale bars are shown on the right. The most salient features of this matrix are that all tools are positively correlated with one another, but also that statistical potentials trained by machine learning (MAESTRO, DUET, NeEMO, PoPMuSiC, mCSM) form a cluster of tools particularly well-correlated with one another. This high correlation suggests this subset of tools could be utilized together with a simple majority vote to improve prediction results [342].

Table 6.3: Performance of stability prediction tools in real-world protein engineering applications

Tool(s) Used	Mutation	Predicted ΔΔG (kcal/mol)	Experimental Results
Reference: Gilis *et al.* [358]			ΔΔG (kcal/mol)
PoPMuSiC1.0	K331F	1.35	1.75 / Success
PoPMuSiC1.0	K331I	1.30	0.84 / Success
PoPMuSiC1.0	K331V	0.80	0.89 / Success
PoPMuSiC1.0	K331T	0.70	-0.54 / Failure
Reference: Cabrita *et al.* [359]			ΔΔG (kcal/mol)
PoPMuSiC1.0	K45F	2.63	0.43 / Success
PoPMuSiC1.0	L56V	1.06	0.00 / Failure
PoPMuSiC1.0	Q58F	1.27	0.60 / Success
PoPMuSiC1.0	E106G	1.64	1.40 / Success
PoPMuSiC1.0	S135G	1.07	0.00 / Failure
Reference: Yang *et al.* [360]			ΔT_m (°C) [ΔΔG (kcal/mol)]
PoPMuSiC1.0	E336C	3.05	3.7 [4.7] / Success
PoPMuSiC1.0	A273C	2.90	0.0 [0.0] / Failure
PoPMuSiC1.0	E400I	2.39	2.2 [2.7] / Success
Reference: Zhang *et al.* [361]			$\Delta t_{1/2}$ (% change)
PoPMuSiC2.0	S92A	0.81	1 / Failure
PoPMuSiC2.0	D93G	1.06	18 / Failure
PoPMuSiC2.0	D174A	0.66	-5 / Failure
PoPMuSiC2.0	S187F	3.23	656 / Success
Reference: Komer *et al.* [362]			ΔT_m (°C) [ΔΔG (kcal/mol)]
FoldX	S57D	0.82	-0.8 [-0.7] / Failure

FoldX	Y60I	3.24	0.1 [0.1] / Failure
FoldX	Y60L	3.12	0.8 [0.7] / Success
FoldX	N93K	1.12	2.1 [1.9] / Success
FoldX	V110L	1.35	-0.9 [-0.8] / Failure
FoldX	T164K	0.95	-0.6 [-0.5] / Failure
FoldX	V404A	1.36	-1.5 [-1.3] / Failure
FoldX	V430F	1.68	1.5 [1.3] / Success
FoldX	N439G	2.59	ND / Failure
FoldX	S13P	1.43	0.7 [0.6] / Success
FoldX	T41V	1.55	-0.3 [-0.3] / Failure
FoldX	A201P	0.91	0.2 [0.2] / Failure
FoldX	S222K	1.43	0.0 [0.0] / Failure
FoldX	T257V	1.01	0.1 [0.1] / Failure
FoldX	T257K	1.17	-2.0 [-1.8] / Failure
FoldX	T273K	0.88	-0.1 [-0.1] / Failure
FoldX	T395P	2.29	ND / Failure
FoldX	T408D	1.22	0.2 [0.2] / Failure
Reference: Silva *et al.*[a] [363]			$\Delta t_{1/2}$ (% change)
PoPMuSiC2.0	G25D	0.62	~0 / Failure
PoPMuSiC2.0	G55P	0.75	~0 / Failure
PoPMuSiC2.0	G55V	0.56	35 / Success
PoPMuSiC2.0	G55S	0.29	~0 / Failure
PoPMuSiC2.0	A67P	0.75	25 / Success
PoPMuSiC2.0	D158L	0.65	~0 / Failure
PoPMuSiC2.0	K243D	0.63	~0 / Failure
PoPMuSiC2.0	K243P	0.75	~0 / Failure
PoPMuSiC2.0	G326A	0.48	~0 / Failure
PoPMuSiC2.0	G326E	0.15	24 / Success
PoPMuSiC2.0	E434P	0.68	~0 / Failure
PoPMuSiC2.0	V541P	0.16	~0 / Failure

PoPMuSiC2.0	G558W	0.56	~0 / Failure
PoPMuSiC2.0	G558D	1.39	~0 / Failure
Reference: Song *et al.* [341]			ΔΔG (kcal/mol)
FoldX + TI	H22K	0.6	0.3 / Failure
FoldX + TI	H22W	2.8	1.2 / Success
FoldX + TI	V25I	2.3	0.4 / Success
FoldX + TI	T30M	5.5	0.5 / Success
FoldX + TI	A33Y	2.2	0.5 / Success
FoldX + TI	T50M	1.8	-0.5 / Failure
FoldX + TI	T54Y	2.6	0.2 / Failure
FoldX + TI	A81M	1.2	-0.4 / Failure
FoldX + TI	V88L	1.1	-0.2 / Failure
FoldX + TI	V90I	1.3	-0.7 / Failure
FoldX + TI	L106M	3.1	ND / Failure
FoldX + TI	N107F	2.3	0.9 / Success
FoldX + TI	N107Y	2.9	0.8 / Success
FoldX + TI	D109E	7.9	0.3 / Failure
FoldX + TI	M111F	2.4	2.9 / Success
FoldX + TI	V120I	1.5	1.4 / Success
FoldX + TI	N124F	3.6	1.1 / Success
FoldX + TI	N124Y	4.2	2.6 / Success
Reference: Wijma *et al.* [364]			ΔT_m (°C) [ΔΔG (kcal/mol)]
FoldX	Q7M	1.9	0.2 [0.0] / Failure
FoldX	R9P	2.4	ND / Failure
FoldX	S12M	2.4	-1.0 [-0.2] / Failure
FoldX	T22D	2.9	ND / Failure
FoldX	A40P	4.5	-1.8 [-0.4] / Failure
FoldX	A41P	1.9	-5.3 [-1.2] / Failure

FoldX	A48F	1.9	ND / Failure
FoldX	E49P	1.9	ND / Failure
FoldX	E68L	2.4	ND / Failure
FoldX	Y96W	1.7	-1.8 [-0.4] / Failure
FoldX	S111M	2.9	-5.0 [-1.1] / Failure
FoldX	G129S	2.9	-2.3 [-0.5] / Failure
FoldX + MD	K13P	0.5	-4.0 [-0.9] / Failure
FoldX + MD	S15P	3.1	1.0 [0.2] / Failure
FoldX + MD	S15K	1.2	-0.3 [-0.1] / Failure
FoldX + MD	A16R	1.0	-0.5 [-0.1] / Failure
FoldX + MD	A20P	3.1	-2.1 [-0.5] / Failure
FoldX + MD	I27M	1.9	-1.8 [-0.4] / Failure
FoldX + MD	D33R	3.1	-2.3 [-0.5] / Failure
FoldX + MD	A41S	1.2	0.3 [0.1] / Failure
FoldX + MD	A66P	1.7	0.8 [0.2] / Failure
FoldX + MD	T85I	4.8	5.8 [1.5] / Success
FoldX + MD	T85L	1.7	2.5 [0.6] / Success
FoldX + MD	T85P	0.7	-6.0 [-1.4] / Failure
FoldX + MD	N92K	2.6	7.3 [1.9] / Success
FoldX + MD	N92R	4.1	1.8 [0.4] / Success
FoldX + MD	Y96F	1.7	2.8 [0.7] / Success
FoldX + MD	L106R	0.5	-2.3 [-0.5] / Failure
FoldX + MD	K110R	1.7	-1.5 [-0.4] / Failure
FoldX + MD	S111K	2.4	-2.3 [-0.5] / Failure
FoldX + MD	S111R	2.9	0.0 [0.0] / Failure
FoldX + MD	A142R	0.2	-2.8 [-0.7] / Failure
Rosetta-ddG + MD	G18N	1.2	0.3 [0.1] / Failure
Rosetta-ddG + MD	E45K	1.9	2.0 [0.5] / Success
Rosetta-ddG + MD	D65E	1.2	0.5 [0.1] / Failure
Rosetta-ddG + MD	D68Q	1.9	0.5 [0.1] / Failure

Rosetta-ddG + MD	A72R	0.2	-1.8 [-0.4] / Failure
Rosetta-ddG + MD	T76A	1.2	0.8 [0.2] / Failure
Rosetta-ddG + MD	T76K	2.4	1.5 [0.4] / Success
Rosetta-ddG + MD	L94F	4.1	0.0 [0.0] / Failure
Rosetta-ddG + MD	Q121I	3.3	ND / Failure
Rosetta-ddG + MD	Q121V	1.7	-10.0 [-2.2] / Failure
Rosetta-ddG + MD	L122F	2.2	-7.5 [-1.7] / Failure
Rosetta-ddG + MD	E124D	1.2	1.3 [0.3] / Failure
Rosetta-ddG + MD	E124K	1.9	-1.3 [-0.3] / Failure
Rosetta-ddG + MD	I127F	2.9	-2.5 [-0.6] / Failure
FoldX + MD / Rosetta-ddG + MD	A19K	1.2 / 1.7	2.5 [0.6] / Success
FoldX + MD / Rosetta-ddG + MD	D24R	2.9 / 3.1	-3.8 [-0.9] / Failure
FoldX + MD / Rosetta-ddG + MD	D33K	2.4 / 2.4	-1.5 [-0.4] / Failure
FoldX + MD / Rosetta-ddG + MD	Y62W	1.2 / 3.8	0.0 [0.0] / Failure
FoldX + MD / Rosetta-ddG + MD	T85V	3.3 / 4.5	6.8 [1.7] / Success
FoldX + MD / Rosetta-ddG + MD	E124R	0.2 / 1.0	-2.8 [-0.7] / Failure
Reference: Deng *et al.* [365]			$\Delta t_{1/2}$ (% change)
PoPMuSiC2.1	K84W	1.83	-56 / Failure
PoPMuSiC2.1	E151I	1.40	-62 / Failure
PoPMuSiC2.1	N302W	1.29	50 / Success
PoPMuSiC2.1	S342F	1.25	-22 / Failure
PoPMuSiC2.1	P477V	1.21	59 / Success
Reference: Larsen *et al.* [366]			$\Delta t_{1/2}$ (% change)
PoPMuSiC2.0	K31P	0.87	55 / Success
PoPMuSiC2.0	N71A	0.57	24 / Success
PoPMuSiC2.0	G116D	0.15	28 / Success
PoPMuSiC2.0	Q171S	0.73	-37 / Failure
PoPMuSiC2.0	G287S	0.63	4 / Failure

Reference: Heselpoth *et al.* [367]			ΔT_m (°C) [$\Delta\Delta G$ (kcal/mol)]
FoldX	Q332H	2.19	0.1 [0.1] / Failure
FoldX / Rosetta-ddG	D330Y	1.09 / 2.51	ND / Failure
FoldX / Rosetta-ddG	Q332V	1.79 / 1.29	-2.0 [-1.5] / Failure
FoldX / Rosetta-ddG	C345T	1.20 / 2.70	-0.1 [-0.1] / Failure
FoldX / Rosetta-ddG	D375Y	2.49 / 1.53	-2.4 [-1.7] / Failure
FoldX / Rosetta-ddG	T381Y	1.32 / 2.65	ND / Failure
FoldX / Rosetta-ddG	V384Y	1.04 / 5.57	0.4 [0.3] / Failure
FoldX / Rosetta-ddG	C404I	2.17 / 5.44	-1.0 [-0.7] / Failure
FoldX / Rosetta-ddG	T406R	1.05 / 1.29	2.3 [1.8] / Success
FoldX / Rosetta-ddG	T421I	2.38 / 3.43	-10.4 [-7.0] / Failure
Reference: this work^{b}			$\Delta\Delta G$ (kcal/mol)
Rosetta-ddG / PoPMuSiC2.0	D49N	0.94 / 0.39	0.2 / Failure
FoldX / Rosetta-ddG / PoPMuSiC2.0	K53V	0.22 / 2.03 / 0.68	0.8 / Success
FoldX / Rosetta-ddG / PoPMuSiC2.0	A62V	1.34 / 0.49 / 0.27	0.5 / Success
FoldX / Rosetta-ddG	E66L	0.84 / 1.26	-0.9 / Failure
FoldX / Rosetta-ddG / PoPMuSiC2.0	E66Y	0.89 / 0.97 / 0.40	-0.6 / Failure
FoldX / Rosetta-ddG / PoPMuSiC2.0	A68G	1.09 / 0.42 / 0.81	0.1 / Failure
FoldX / Rosetta-ddG / PoPMuSiC2.0	Q78I	0.58 / 1.79 / 1.42	-2.8 / Failure
FoldX / Rosetta-ddG / PoPMuSiC2.0	D85P	2.31 / 0.31 / 0.41	0.1 / Failure
FoldX / Rosetta-ddG	R90L	0.95 / 1.19	0.0 / Failure
FoldX / Rosetta-ddG	D93P	1.16 / 0.43	0.0 / Failure
Overall Summary (shown in Table 6.1)			
Tool Used	# of mutations tested / # of reports	Success Rate (PPV)	Average change in stability ($\Delta\Delta G_{exp,avg}$)
FoldX	49 / 4	14%	-0.42 ± 1.40
FoldX + MD	26 / 1	27%	0.0 ± 0.79

Rosetta-ddG	19 / 2	16%	-0.75 ± 1.87
Rosetta-ddG + MD	20 / 1	20%	-0.17 ± 0.83
PoPMuSiC	47 / 8	43%	0.53 ± 1.48
FoldX + TI	18 / 1	56%	0.66 ± 0.96
All	**148 / 12**[c]	**30%**	**0.03 ± 1.25**

Success or failure for each tool is tabulated individually. Thus if both FoldX and Rosetta-ddG predicted a mutation to be stabilizing but experimental tested showed it to be a failure, this is counted as a failure for both FoldX and Rosetta-ddG individually. In the case of the addition of MD or TI, these are counted as separate cases because they were used as a second-tier filter and thus they represent something above and beyond just a single method.

Depending on the source, three possible measures of change in protein stability are reported: change in thermdynamic stability ($\Delta\Delta$G), change in melting temperature (ΔT_m), and change in thermal inactivation half-life ($\Delta t_{1/2}$). Because the tools themselves make predictions in terms of $\Delta\Delta$G, this measure is the most useful. Fortunately, $\Delta\Delta$G can be estimated from ΔT_m following the method of Rees and Robertson [343]. This estimated $\Delta\Delta$G is reported in square brackets in the table. A mutation with a reported or estimated (as above) $\Delta\Delta$G is considered a success if it stabilizes the protein by more than 0.3 kcal/mol. For a $\Delta\Delta$G estimated from ΔT_m this corresponds to a change of $\sim$1 °C, though the exact value varies depending on protein size and the T_m of the WT [343]. There is no reported method of estimating $\Delta\Delta$G from $\Delta t_{1/2}$. Thus, only mutations that had $\Delta\Delta$G or ΔT_m measured are counted for the average change in stability reported at the end of the table. However, mutations with measured $\Delta t_{1/2}$ are still tabulated as success or failure based on an increase in $t_{1/2}$ of at least 20%, a value suggested by the work of Silva *et al.* [363].

MD refers to the use of molecular dynamics and human interpretation of MD results to screen out mutations recommended by the automated tools. This does not include more sophisticated methods like thermodynamic integration (TI).

ND indicates the mutant could not be experimentally tested for reasons such as poor expression or solubility, and as such is tabulated here as a failure.

[a] Improvements in $\Delta t_{1/2}$ that were reported as negligible from an initial screen in the source citation, are tabulated here as failures with approximately no change in thermal inactivation half-life ($t_{1/2}$).

[b] 11 tools were used to decide on each mutation. For simplicity and comparison to the other studies, only predictions from FoldX, Rosetta-ddG, and PoPMuSiC predictions are shown.

[c] The total is less than the sum of the individual tools because some mutations were chosen by multiple tools.

Table 6.4: ThreeFoil mutant kinetics and stability analysis

ThreeFoil	ln(k_f)	ln(k_u)	m_f	m_u	C_{mid}	ΔΔG
Mutant	(sec^{-1})	(sec^{-1})	(kcal mol^{-1}M^{-1})	(kcal mol^{-1}M^{-1})	(M)	(kcal/mol)
WT	-9.56	-22.0	-6.38	3.20	0.78	0.0
D49N	-9.09	-21.9	-6.32	3.23	0.80	0.2
K53V	-8.03	-21.9	-7.40	3.23	0.78	0.8
A62V	-8.65	-22.0	-6.80	3.25	0.79	0.5
E66L	-9.58	-20.6	-5.61	3.02	0.76	-0.9
E66Y	-9.73	-21.2	-6.80	3.09	0.69	-0.6
A68G	-9.59	-22.1	-6.14	3.29	0.79	0.1
Q78I[a]	*-8.43*[a]	-16.3	ND[a]	2.70	*0.49*[a]	*-2.8*[a]
D85P	-9.05	-21.9	-6.62	3.20	0.78	0.2
R90L	-9.98	-22.4	-5.90	3.33	0.80	0.0
D93P	-9.44	-21.9	-6.32	3.13	0.78	0.0
Multi-mutant[b] “Seq1”	-4.32	-21.4	-6.50	2.71	1.11	**2.8**
Multi-mutant[b] “Seq2”	-3.16	-20.4	-6.74	2.50	1.11	**2.8**

ΔΔG values in bold represent improvements in stability beyond typical experimental errors (0.3 kcal/mol, see work by Pokala et. al [310]) and therefore likely to be real improvements.

[a] Q78I had dramatically reduced solubility which prevented determination of the refolding branch. Refolding in 0.02M GuSCN was used with the WT m_f to estimate k_f and therefore ΔΔG for this mutant.

[b] Multi-mutants were designed to take advantage of ThreeFoil’s three-fold structural and sequence symmetry, and therefore have mutations made at all 3 symmetric positions to amplify the stabilizing effect. “Seq1” is K6V/A15V/D38P/K53V/A62V/D85P/K100V/A109V/D132P, while “Seq2” is D2N/K6V/A15V/D38P/D49N/K53V/A62V/D85P/D96N/K100V/A109V/D132P

Chapter 7

Molecular Dynamics of Ligand Binding: Equilibrium, Umbrella Sampling, and Metadynamics Simulations

7.1 Context

Throughout my graduate studies I collaborated on several projects by providing molecular dynamics expertise. Several of these projects have been published, but many have not, either because of low confidence in the results and the computational/simulation time that would have been required to gain this confidence, or difficulties comparing to experimental data. The purpose of this chapter is to briefly highlight these results with a focus on which approaches worked and which did not. For the latter case, I attempt to provide some basis for understanding what went wrong and how that may be fixed or recognized more easily in the future.

7.2 Summary

Molecular dynamics, both as simple equilibrium simulations and more advanced biased simulations such as umbrella sampling and metadynamics, were used to interrogate the behaviour of several protein-ligand systems and a very simple water-dimer system. Equilibrium simulations of streptavidin and biotin in the gas phase were successful in understanding the kinetic stability of these interactions within a mass spectrometer. The use of umbrella sampling to quantify the energetics and ligand exit-pathways of two different protein-ligand systems, also in the gas phase, gave results that were either in poor agreement with experiment or that needed considerable additional resources to be compared properly. In both cases this failure is attributable to high kinetic barriers in the gas phase which make proper equilibration challenging. By contrast, both equilibrium and umbrella sampling simulations of a protein and pseudo-ligand in explicit solvent were fairly successful, giving results broadly inline with experiments (though some issues with equilibration persist here as well). Unfortunately, a lack of one to one mapping between the experimental and simulation information stymied efforts to use these simulations effectively and be confident in their results (though recent developments could revitalize such efforts).

Finally, a very simple case of two water molecules in the gas phase was studied using metadynamics in order to provide a somewhat rough, but accurate energy surface from which to test fitting methods for approximating the long-range portion of any such binding event landscape. Overall the results are mixed. The use of gas-phase simulations (no solvent) on large protein-ligand systems should be approached with considerable care, and while similar systems in solvent can provide valuable results, determining proper equilibration of the simulations is a challenging task. Moreover, while molecular dynamics simulations can provide a vast wealth of information, computational chemists should put considerable forethought into ensuring their results can be compared directly to some kind of experimental data in order to have confidence in other results or interpretation gleaned from those simulations.

7.3 Introduction

7.3.1 Molecular dynamics

Molecular dynamics (MD) is a computational technique that simulates how atoms move and interact. Given the extremely complex equations that govern atomic interactions — quantum mechanics (QM) — providing a complete and accurate solution at the level of a whole protein and ligand solvable in water is beyond current computational capabilities. An alternative is the use of molecular mechanics (MM), where atomic interactions are represented with simplified equations. For instance, van der Waals interactions are modelled by a Lennard-Jones potential [368], covalent bonds to hydrogens are often considered rigid both in length and angle [369], electrostatic interactions are only explicitly calculated within a certain cutoff [370] and other approximations are used to simplify a mathematical simulation of the system. These equations are used go compute the forces on each atom, and integration of the classical equations of motion (i.e. Newtonian mechanics) then provides a simulation of the system given these numerous and varied simplifications. Where chemistry in important, QM may be used for a very small subset of the system (e.g. the active site of a protein), while MM is used for the remainder of the protein [371, 372]. This is still extremely expensive and limits the timescale of simulations. Thus, in cases where chemistry is not critical (such as ligand binding), pure MM (often simply called MD) can be used, as is the case in this chapter. Despite the approximations used in MD, it can often produce results in excellent agreement with experiment [373, 374].

In the simplest case, an MD simulation begins with the system in a known configuration (e.g. from an X-ray structure), and is allowed to evolve according to the aforementioned physical rules under a dynamic influence like temperature. If the system is unstable (a folded protein at high temperature), then MD may reveal physically meaningful details of how that protein unfolds, such as which regions are the least stable [375]. If the system is stable, for instance a ligand sitting in a protein's binding pocket, the simulation may reveal details of the structural/configurational ensemble of the bound state. However, because simulation timescales are typically much shorter — ns to µs — than those of interesting biomolecular processes (e.g. protein folding from the denatured state, or ligand binding from solution) — ms to sec — observing a biologically relevant event is rare (though special-purpose computers are making considerable headway here [376]). One obvious solution to this problem is to raise the simulation temperature, thereby flattening the energy landscape and making transition over energy barriers more feasible. Of course, properties thus obtained may not reflect the biologically relevant temperatures. There exist solutions to this problem, such as parallel tempering (replica exchange) or simulated tempering, and other Hamiltonian-exchange based approaches [377]. Such methods, however, suffer from the problem of protein unfolding at high temperature, and are therefore limited in the temperature range (and thus utility) they can provide. While restraints can be placed on the protein structure to avoid the aforementioned problems, these may inhibit exactly the kind of motions the high temperature where meant to generate (i.e. protein dynamics enabling ligand binding). Thus, while these methods have been applied to many problems with excellent results, and continue to be developed, they require considerable forethought and expert consideration for each individual system. Because MD is a simulation, however, we are not limited to simply observing a mock-up of reality, but can interact with and manipulate it. In particular, the simulation can be steered by non-physical or "external" forces so as to more rapidly reveal the information we are interested in (e.g. the ligand can be “pushed” out of the binding site to reveal the lowest energy exit-pathway). These techniques are usually referred to as biased-methods, and two such methods, umbrella sampling, and metadynamics, are used in this chapter.

7.3.2 Umbrella sampling

One of the oldest methods for biasing or steering a simulation is known as umbrella sampling. I leave a detailed description of the underlying statistical mechanics and formulation of the method to others [378, 379], and here provide a brief description of the approach. The overall goal of umbrella sampling is to construct a free energy surface (FES) for a particular process occurring along a particular reaction coordinate. For instance, if we imagine a ligand in its protein binding site, the reaction coordinate may be the distance from the center of mass of the ligand to the center of mass of the binding pocket. In theory, the FES could simply be generated from the population/probability distribution across an unbiased simulation. That is, the distance of interest, or reaction coordinate is recorded at a particular time interval and this is used to build up a distribution which can be converted to a free energy surface using the Boltzmann distribution as:

$$E_i = k_B T * \ln p_i \tag{7.1}$$

where, p_i is the probability of finding the system at reaction coordinate i, E_i is the energy of the system at reaction coordinate i, k_B is the Boltzmann constant, and T is the simulation temperature.

As noted earlier, however, simulation times are frequently too short to capture even a single binding or unbinding event, let alone enough to generate a reliable distribution from which to calculate a FES. Since evolution of the system over time is defined by a large collection of forces, we can simply add in an additional force with little impact on the simulation’s computational performance. For umbrella sampling this is typically a harmonic restraint which produces a spring-like force restraining the ligand near a particular position along the reaction coordinate:

$$F = k(x - x_{eq}) \tag{7.2}$$

where, F is the force, k is the spring constant, x is current position along the reaction coordinate and x_{eq} is the equilibrium position of the restraint.

By running numerous simulations each with a different x_{eq} the full reaction coordinate can be explored. Of course, any FES built from these simulations will reflect the simulation with the harmonic restraint and thus be unrealistic. But, the potential energy of the harmonic restraint is precisely known from k and x_{eq} as the integral of the force:

$$U = \frac{1}{2}k(x - x_{eq})^2 \tag{7.3}$$

Thus, the potential energy of the spring can be subtracted from each simulation's FES to reconstruct the underlying (internal or "true") FES in the vicinity of (x_{eq}). How much of this underlying FES is reliable will depend on the extent to which each simulation explored the reaction coordinate. This exploration will depend on both the underlying true free energy surface (a more rugged surface impeding exploration) and the strength of the harmonic restraint (a stronger restraint impeding exploration). Nevertheless, these small local fragments of the underlying FES can then be "stitched" together to reconstruct the full FES. While simply merging fragments so as to minimize the deviation in overlapping regions can work in simple cases, the more rigorous weighted histogram analysis method (WHAM) is typically used [380]. The constructed FES can then be used to calculate quantities like the binding free energy and rate constants. It should be noted that while the description presented uses individual simulation windows each with their own restraint potential, and this method is used in this chapter, it is possible to do all of this within a single longer simulation. Given the plateau of single processor power seen recently, however, any ability to make the computations more parallel is welcome.

7.3.3 Metadynamics

There have been many attempts to develop improved biasing methods since the inception of umbrella sampling. One of the most popular, metadynamics, works in what might be conceptually considered the opposite manner [381, 382]. That is, rather than driving the system towards a specific reaction coordinate, small repulsive potentials or "hills" are added to the simulation in a time- or history-dependant manner in order to promote exploration of the whole reaction coordinate. These repulsive "hills", which usually take a Gaussian form (see equation 7.4 below), act by "flooding" or "elevating" the local FES in order to make the current position on the reaction coordinate higher energy and less favourable.

$$U = a \exp \frac{-(x - b)^2}{2c^2} \tag{7.4}$$

where a is the height or strength of the hill (i.e. potential energy in kcal/mol), b is the position (placed at the simulation's current position along the reaction coordinate to encourage exploration), and c controls the width (i.e. how close to this position on the reaction coordinate the system will have to come before feeling this repulsion).

Eventually, the underlying FES will be completely filled with these repulsive hills, and the ligand will appear to move freely along the reaction coordinate (e.g. to and from the binding site) as though it were on a completely flat FES with no minima or barriers. If this occurs, the simulation can be stopped and the sum total of the many hills which have been added (the full biasing potential) can be subtracted from the aforementioned flat free energy surface to yield the underlying FES (though in practice the sign of the full biasing potential is often just reversed to generate an estimate of the underlying FES).

Thus, metadynamics obtains the same final information as umbrella sampling, but has the advantage that all of this can be easily done in a single simulation without need for reconstruction techniques (the previously mentioned stitching together of small FES fragment). This is particularly valuable when constructing a reaction coordinate along multiple dimensions (for instance distance from the ligand binding site and orientation of the ligand), where reconstruction by umbrella sampling becomes increasingly difficult and unreliable. A critical problem with metadynamics (and all biased-methods including umbrella sampling) is seen when there is an unbiased degree of freedom that limits exploration along the reaction coordinate. For instance, if the reaction coordinate is the distance from the binding pocket, but the ligand can only bind in a particular orientation, then the orientation is an unbiased degree of freedom. In this case, if the ligand approaches the binding site in the incorrect orientation it will be unable to bind. While this is a general problem, the result in metadynamics is particularly undesirable. In this case many repulsive hills will continue to be added trying to force the ligand into the unexplored binding site, thereby greatly overestimating the free energy of the unbound state. Eventually the ligand may enter the binding site, but if these same slow degrees of freedom prevent (or others) then prevent leaving, the bias will not only fill up appropriately, but subsequently overfill. This process can continue to cycle, with the biased potential appearing to have a hysteresis. Eventually this may generate unphysical conditions, such as the ligand being forced inside the center of the protein, and thereby unfolding it. Thus, metadynamics suffers from problems with proper convergence of the bias. If the slow degrees of freedom are known, they can be biased against, but this is rarely something that is simple to identify. Some solutions to this problem have been developed, such as well-tempered metadynamics [383] and adaptive biasing force [384], but these are not flawless.

In this chapter umbrella sampling and standard metadynamics are both used along with equilibrium simulations with varying degrees of success. As is usually the case, failure cannot be blamed on the methods, but rather on their implementation and perhaps naive expectations of the user (the author in this case).

7.4 Results

7.4.1 Equilibrium simulations of a protein-ligand complex in the gas phase

Examining protein-ligand binding experimentally can be challenging if there isn't a clear signal produced when the ligand binds. Often changes in intrinsic fluorescence or circular dichroism of the protein or the ligand are used. For example in Chapters 2 and 5 changes in intrinsic fluorescence of ThreeFoil was used to detect binding of the carbohydrate ligand. But such changes are not universally present for all systems. Mass spectrometry, by contrast, is more broadly applicable (all ligands have mass after all).

Mass spectrometry requires the sample to be ionized into a gaseous phase just prior to analysis. Thus, while it may seem reasonable to use mass spectrometry to quantify protein-ligand binding interactions in the gas phase itself, one may be left wondering how well this translates to the biologically relevant solution phase? Work by the Klassen group [385] showed that in fact, the ligand-bound form can be preserved upon ionization and thus the fraction of ligand bound protein found in mass spectrometry is equivalent to that in solution.

There was interest within the Klassen group to determine if the specific protein-ligand interactions present in solution are also maintained the gas-phase. To address this question I collaborated with the Klassen group and used equilibrium MD to determine if there are significant changes in the protein-ligand interactions of the streptavidin-biotin system when moving from solution to the gas phase.

A significant problem encountered in performing simulations meant to mimic conditions within the mass spectrometer, is to account for ionization of the protein. Specifically, the protein will acquire an unnatural (at least compared to the solution phase) charge state, and the location of these charges needs to be exactly defined during MD. In this case, the streptavidin tetramer was known to have a net charge of +12 during mass spectrometry. Assuming an equal charge distribution across each monomer, there are 3 positive charges to assign to specific residues. As positive charges are most likely to be on lysine, arginine, or histidine and there are 9 such residues per monomer, this gives 84 possible charge configurations. From these 84 possibilities, 15 were suggested as reasonable based on visual inspection by members of the Klassen group. Notably, in all 15 cases Lysine 121 was positively charged.

Each of the 15 configurations was energy minimized, simulated by MD until equilibrium was reached, and then simulated for production and data collection (see Methods). Analysis was performed to determine which protein residues were interacting with the biotin ligand, in terms of both van der Waals contacts and hydrogen bonds (see Methods). The results of this analysis are compared against those in solution taken from the crystal structure (Figure 7.1). Most notably, only one interaction differs: a hydrogen bond between Serine 88 and the acid tail of biotin is replaced by a hydrogen bond from Lysine 121 of an adjacent monomer (Figure 7.2).

Since the initial coordinates for simulation were taken from the crystal structure (see Methods) this simulation is analogous to suddenly ionizing the protein-ligand complex, and the minimal change in interactions after equilibrating the simulation suggests that specific interactions native to the solution structure remain intact in the gas phase. It is noteworthy that the only interaction that differed from the solution structure to the gas-phase simulations involved Lysine 121, which had been consistently chosen as charged in all 15 simulated configurations. It is possible this new interaction is largely an artifact of that choice and perhaps a less subjective method of choosing the charge states would yield a result in complete agreement with the crystal structure. On the other hand, the total simulation time of the system was 20 ns. It could be argued that this is an insufficiently long simulation to capture the structural rearrangements that might occur. In particular, while measures such as structural root mean squared deviation (RMSD), and potential energy suggested the simulation had reached equilibrium, it is impossible to know if this is the same equilibrium that would be reached had the simulation been run for ms to reach the time-scale of the experiment. At the time of this study simulations of protein systems typically ranged from 1 to 10 ns and this was largely the basis for selecting this time. Given recent hardware advances such as general purpose graphical processing units (GPGPUs) from nVidia (http://www.nvidia.ca/object/tesla-supercomputing-solutions.html) or special purpose computers like Anton from DESRES [386], it may be possible in the near future to examine this question directly. Though it should also be noted that modern protein forcefields are parameterized for use with explicit solvent and it is not clear how they may behave during a long simulation.

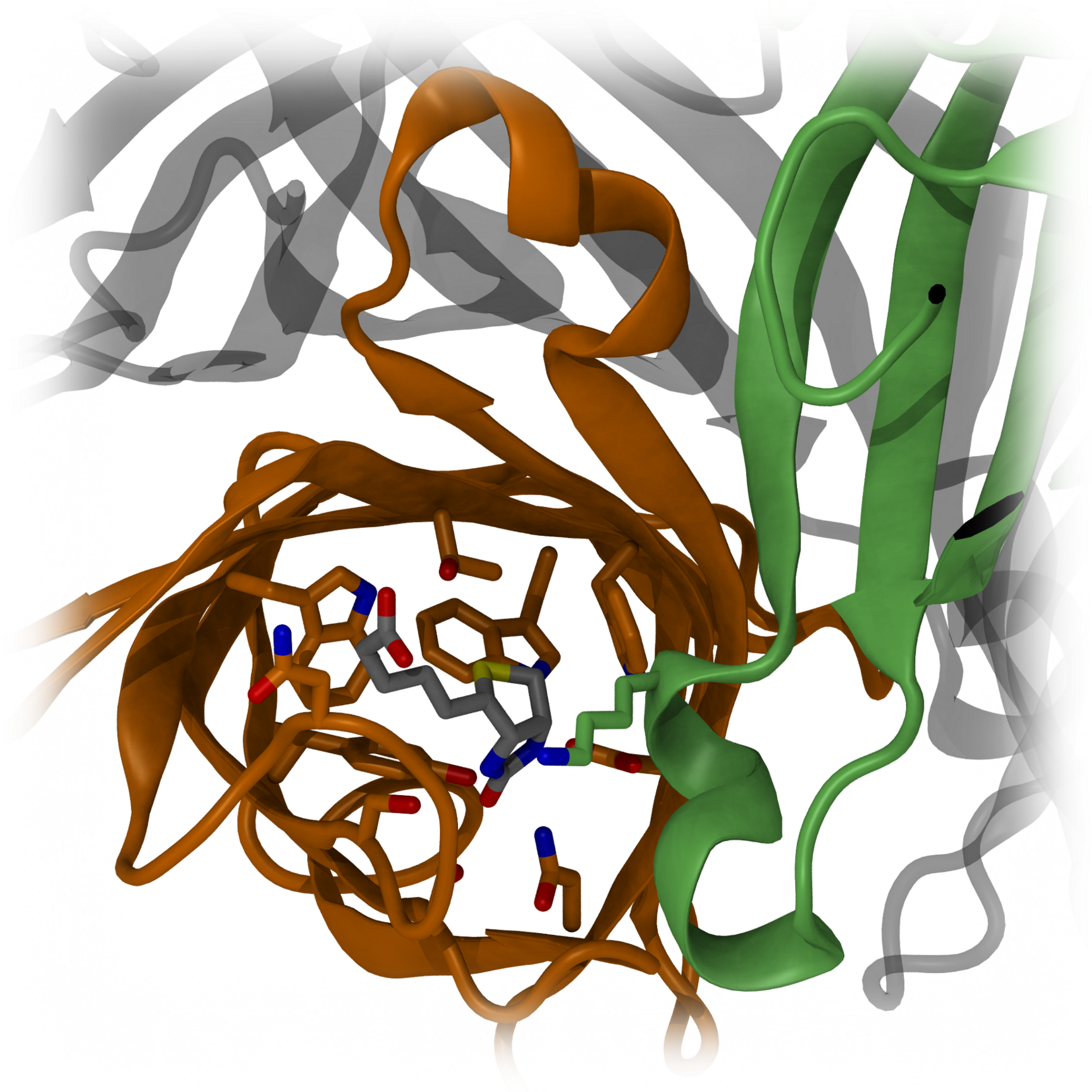


Figure 7.1: The streptavidin-biotin complex in solution. A single streptavidin monomer is shown as an orange cartoon, the adjacent monomer as a green cartoon, and the remaining two as ghosted cartoons. The biotin molecule bound to the orange monomer is shown as sticks with grey carbon, while the remaining biotin molecules are hidden. Sidechains interacting with the biotin molecule (see Figure 7.2) are shown as sticks with carbon colored to match the monomer they are from.

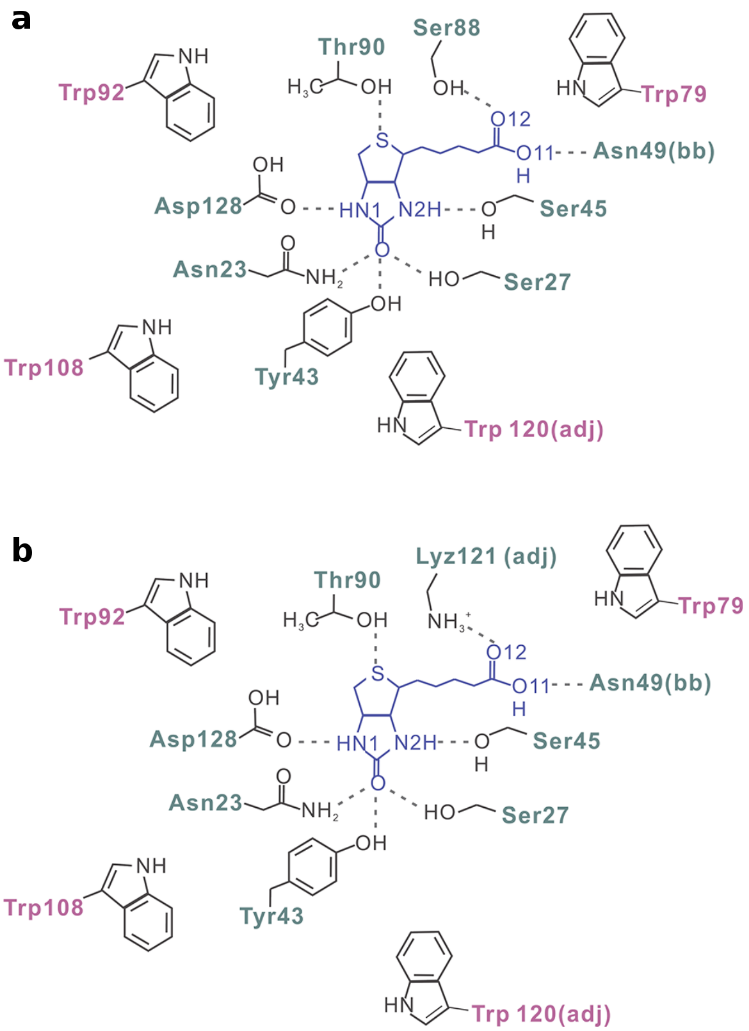


Figure 7.2: Solution *versus* gas-phase interaction maps of streptavidin-biotin. Interaction maps for **(a)** the WT (S4+4B) complex obtained from the crystal structure and **(b)** the WT $(S_4{+}B)^{12+}$ ion determined from MD simulations performed using 15 different charge configurations. Biotin is shown at the center in blue, side-chain atoms shown in black. Hydrogen bonded interactions are shown in black dashed lines indicating the partners. Residue names and numbers are shown in green hydrogen bonds and magenta for van der Waals. Interactions with the backbone rather than sidechain are indicated with "bb" and residues from adjacent streptavidin subunits are indicated with "adj".

7.4.2 Umbrella sampling of protein-ligand complexes in the gas phase

The Klassen group had used mass spectrometry to analyze both energetics and activation energies for several other protein-ligand systems beyond the streptavidin-biotin complex. β-lactoglobulin is a fatty acid binding protein and mass spectrometry had been used to examine the relationship between fatty acid chain length and binding/activation energetics [387] (Figure 7.4a). The single chain fragment of a monoclonal antibody (scFv) with its trisaccharide ligand Gal-Abe-Man [388] (Figure 7.4b) has been studied with the goal of elucidating interactions between the protein and ligand using mutagenesis. Specifically, experiments demonstrated a specific interaction between histidine 228 and an -OH group of the mannose glycan residue. We wanted to see if MD could be used to recapitulate the free energy surfaces and thus, the activation energies for binding in both of these cases, and subsequently, if we could analyze the trajectories to understand the pathway the ligands take when leaving the binding site. I attempted to use umbrella sampling to answer these questions for the scFv system, while providing MD automation scripts and suggestions to a student, Nobar Jalili, who worked on the β-lactoglobulin system.

A key problem I wanted to solve in working on another gas-phase system was the issue of charge configuration. In particular, umbrella sampling requires many simulations to be performed on the same system, potentially for fairly long times, so simulating many configurations as in the previous example wouldn't be possible. Moreover, in these new cases, the number of possible configurations to test was extreme. For instance, in the case of β-lactoglobulin, the protein had a charge of -7 after ionization and 27 potential sidechains to consider, thus giving a total of 888,030 possible charge configurations. To address this issue I wrote a small program which would load the WT structure and AMBER forcefield and then iterate over all possible combinations to find the one with the lowest electrostatic potential energy (see Methods). While this program was used to determine the ideal configuration for β-lactoglobulin, I later wrote a faster but less accurate script for the scFv case, where only the position of the charged groups themselves were considered in calculating the electrostatic potential energy (ignoring all interactions with polar atoms, see Methods).

7.4.3 Accurate and reliable umbrella sampling

Now that a single charge configuration could be objectively decided on I wanted to ensure that the results of umbrella sampling would be reliable. A significant problem in umbrella sampling or any biasing method, and in fact MD in general, is not knowing when a system has truly reached equilibrium and the data can be used for analysis. In umbrella sampling this can occur for any given simulation or “window” and will manifest as that particular FES fragment being incorrect. Because the full FES is made by joining together many fragments the error from any single fragment will propagate to the rest of the surface and aggregate values such as binding energy or activation energy will be poorly defined (this is true even when using the more rigorous WHAM to generate the FES). This is a particularly insidious problem because it is possible, and in fact common, for the simulation to be trapped in a local minimum and therefore appear to most common analyses as though it had reached equilibrium. While there is no fool-proof solution to the problem of equilibration, one very useful, though computationally expensive method, is to perform two simulations for each window, with slightly different initial configurations. Then, in this “sister” method, if the two “sister” simulations reach the same apparent equilibrium there is increased confidence that this is a true equilibrium (i.e global minimum rather than local one). I implemented various automated scripts using python for running the NAMD simulation software (http://www.ks.uiuc.edu/Research/namd/), and subsequently analyzing the results to determine when the above “sister” equilibrium had been reached. A description of the process is given below.

In order to judge if simulations had reached equilibration during umbrella sampling two different starting coordinates are used for each set of umbrella sampling simulations. One from 20 ns after initial equilibration, and one from 40 ns after initial equilibration. Each set of starting coordinates was used to produce a complete set of umbrella sampling simulations, with all bias parameters identical. In addition to the different starting coordinates, the random number seed used for Langevin dynamics was also different for each “sister”. For each window, 10 ns of simulation was performed for each sister, followed by a test for equilibration. First, each sister was tested internally to ensure equilibration within that 10 ns block. This was done by splitting the simulation into equal halves (the first 5 ns and the second 5 ns) and then calculating the mean and standard deviation for each. If the difference between the two means was $< 10\%$ of the largest of the two standard deviations, then that individual simulation was considered to potentially be at equilibrium. If both sisters were judged internally to have potentially reached equilibrium then the two were compared against one another using the same method. That is, the mean and standard deviation was determined for each of the 10 ns sister simulations, and if the means differed by $< 10\%$ of the largest of the two standard deviations, then it was accepted that equilibration really had been reached (though this is really just an improvement in confidence and we can never know for certain). If equilibration had been reached, those simulations were stored for later analysis, otherwise another 10 ns was performed until the test for equilibration passed. An example of this analysis for two failed and one successful case is shown in Figure 7.3.

Final construction of the FES for each system used the final equilibrated 10 ns from each sister, where trajectory data had been collected every 2 ps, thus giving 5000 data points per sister. The FES was constructed using WHAM, implemented in a fast an easily usable manner by Alan Grossfield (http://membrane.urmc.rochester.edu/content/wham). The FES must then be adjusted to account for the fact that the simulation recapitulates an unnatural system of a protein and ligand within an infinitely large vacuum [389, 390]. This is done so that values derived from the FES can be compared to standard conditions such as 1 M ligand.

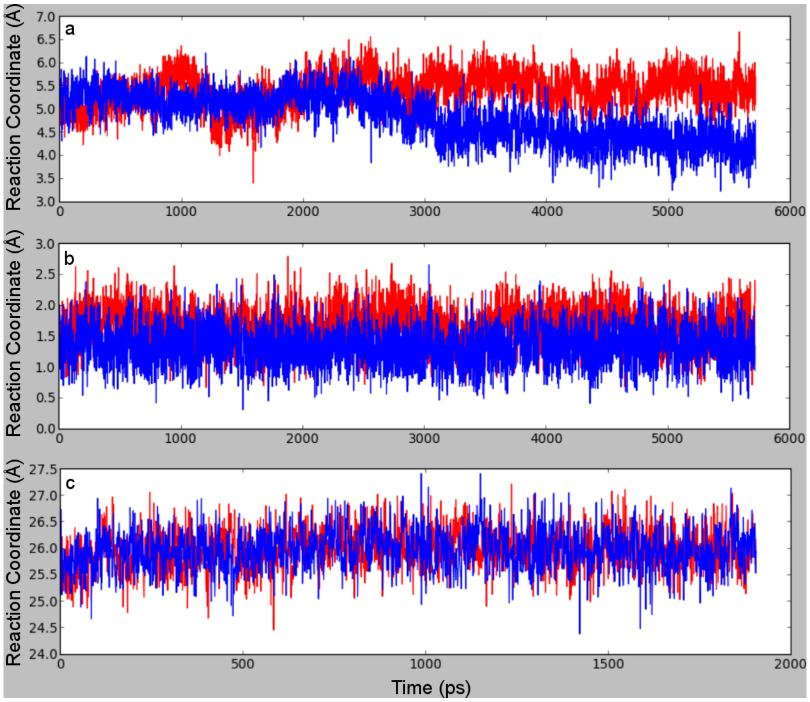


Figure 7.3: Testing for equilibration/convergence in umbrella sampling. Example output from the python code written to test for equilibration/convergence of "sister" umbrella sampling windows is shown. The value of the reaction coordinate over time is shown as the criteria for judging convergence, though structural RMSD or potential energy can also be used and when tested provide the same results. **(a)** Neither of the "sister" simulations appear equilibrated even when considered alone (internally equilibrated) and thus more simulation time is needed. **(b)** Both "sister" simulations are internally equilibrated, and without the use of two simulations these umbrella sampling windows would be considered at equilibrium and used for analysis (typically seen in the literature). However, comparison of the two distributions clearly reveals the red distribution if at a higher value on average than the blue and this difference is beyond what would be expected by chance. **(c)** Both simulations appear internally to be at equilibrium and comparison against one another reveals this is true. These simulations can be used for analysis by WHAM to help reconstruct the FES.

7.4.4 Umbrella sampling of β-lactoglobulin with fatty acid ligands

The free energy surfaces for β-lactoglobulin and fatty acids of differing length can be found in Nobar Jalili's MSc thesis [389]. The main problem that became evident during her simulations was extremely slow equilibration. Despite hundreds of nanoseconds of simulation time for individual umbrella sampling windows, a substantial number of "sister" simulations did not equilibrate. It is impossible to know for how much longer the simulations might have needed to be continued until they reached equilibrium. Constructing the FES for each condition even without equilibrium being reached can at least give some estimate of the expected FES. In this case, however, the predicted binding and activation energies have no correlation to experiment, suggesting these may be very far from equilibrium [389]. Examining the simulations, Nobar suggested the problem arose from long-lived electrostatic interactions between the charged fatty acid head-group and ionized sidechains which were creating kinetic traps. While electrostatic interactions are generally considered to be no stronger than van der Walls or hydrogen bonding in protein-ligand interactions, such considerations apply to the solution phase, where water not only screens electrostatic interactions, but thermal motion allows water molecules to break existing interactions and take their place. Thus, it may be that a gas-phase system is simply not amenable to methods like umbrella sampling which require a transition out of the binding site. It could also be that the single charge configuration selected by the method of minimizing electrostatic potential energy, gave an unrealistic result and alternative configurations would have had ionized groups farther from the head of the fatty acid ligand.

7.4.5 Umbrella sampling of scFv with the Gal-Abe-Man trisaccharide

In the case of the scFv-trisaccharide system (Figure 7.4b, and Figure 7.5), there is no charged group to cause problems with extremely long equilibration times. In fact, equilibration by the stringent "sister" method showed less than 10% of the windows had not reached equilibration, but would have appeared equilibrated by standard single simulation analysis (and thus would be published without any problems being known). Therefore, the results can be interpreted with as much confidence as many reported umbrella sampling simulations, but this lack of complete equilibration should be kept in mind. In order to compute accurate activation energies, which was desired for comparison of kinetics to experiment, the FES need to be generated at several different temperatures to allow the use of transition state theory and the Arrhenius equation [391]. This had been the goal for the β-lactoglobulin system as well, but as the simulations performed at a single temperature showed no correlation with experiment (activation energies which were roughly predicted were clearly in the wrong rank-order [389]) the remaining temperatures were never simulated. In the case of the scFv system, the rank order of binding and activation energies matches very well with experiment. In particular it was found experimentally that mutating histidine 228 to alanine reduced binding and activation energy by the same amount as using an altered ligand in which the mannose was missing an -OH group (deoxy-Mannose), thereby suggesting the interaction of histidine 228 with this -OH group [388]. Notably, simultaneous mutation of histidine 228 and use of the deoxy ligand resulted in the same loss of binding and activation energy, confirming that these two interacted exclusively [388]. The rank-order of binding and activation energies from simulation, which can be estimated from the depth of the binding minima and height of the barrier that minima (Figure 7.5), agrees with the experimental results as the WT FES shows the deepest bound energy minimum and largest energy barrier to long distances form the binding site (unbound), whereas both the protein and ligand mutants and double-mutant are approximately the same.

Nevertheless, as the simulations were performed at elevated temperatures (390 K), the resulting FESs are very shallow and the noise/error (which is typical) has a dramatic impact, making fitting for determination of rate constants and eventually an accurate activation energy, challenging. At high temperature, however, equilibration is expected to faster and this is why such temperatures were initially used (having seen the problems encountered by Nobar with the β-lactoglobulin system). Thus, while it might have been possible to generate accurate FESs usable for fitting at lower temperatures, the simulation time required to this point had already been substantial and the project has been abandoned in lieu of a more efficient technique.

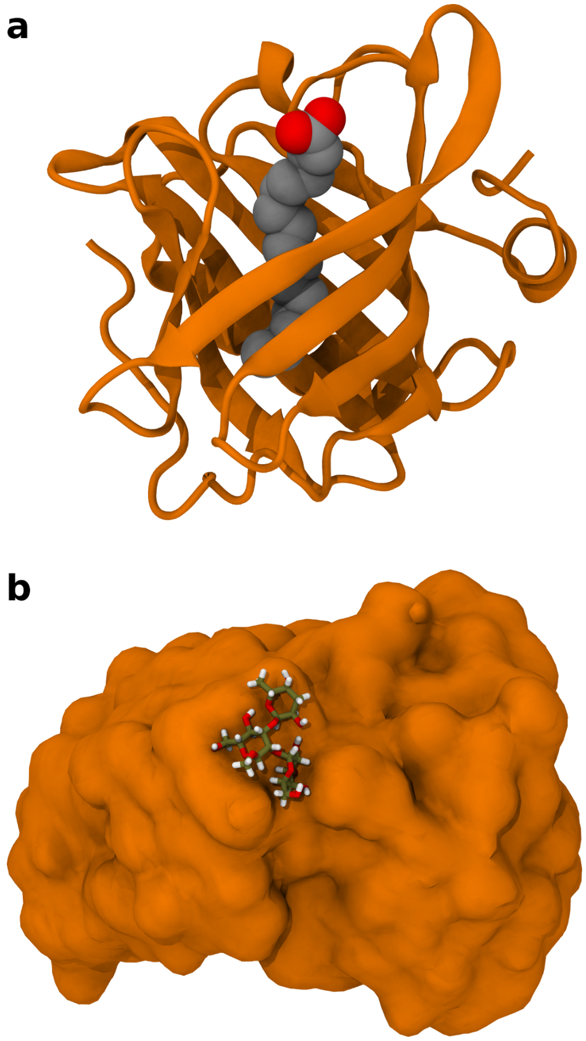


Figure 7.4: β-lactoglobulin with palmitic acid and scFv with the Gal-Abe-Man trisaccharide. **(a)** β-lactoglobulin (as an orange cartoon) with a palmitic acid (as spheres) bound in the core of the protein. **(b)** The scFv monoclonal antibody fragment (as an orange surface) with its trisaccharide ligand (Gal-Abe-Man) as sticks.

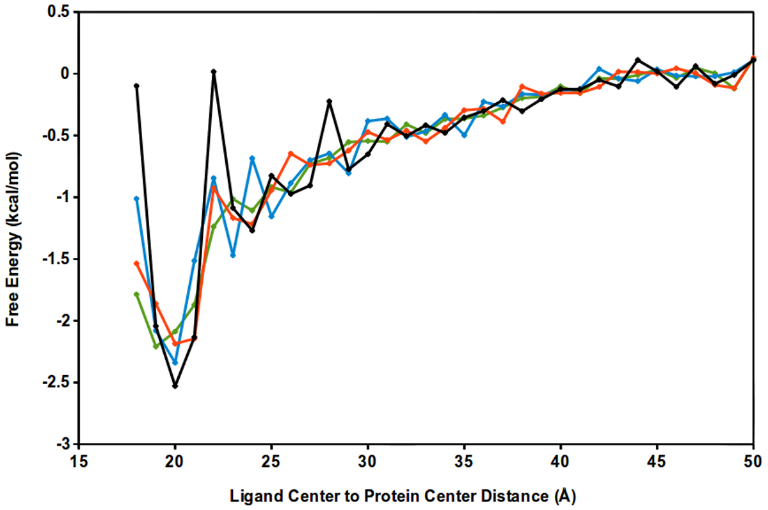


Figure 7.5: FES of scFv and mutants with the Gal-Abe-Man and Gal-Abe-Deoxy-Man trisaccharide. The work needed to escape the binding pocket is shown for WT scFv with the native Gal-Abe-Man trisaccharide (black), the WT scFv with Gal-Abe-deoxyMan (orange), mutant scFv-H228A with Gal-Abe-Man (blue) and mutant scFv-H228A with Gal-Abe-deoxyMan (green). While the WT protein with Gal-Abe-Man has the tightest binding as expected from experiment, the activation energies for all species are at least an order of magnitude too low.

7.4.6 Equilibrium and umbrella sampling of a protein-ligand complex in explicit solvent

Hisactophilin is a small (∼120 amino acid) protein that is N-terminally myristoylated. Myristoylation is a common protein modification that involves the covalent linkage of myristic acid, a C_{14} fatty acid. Often, the myristoyl group switches between a sequestered and accessible state, changing the ability of the protein to bind to the cell membrane or other proteins [392]. In the case of hisactophilin, the myristoyl group switches from sequestered to accessible as the pH moves from basic to acidic, with the switch centered at a pH of 6.95 [249]. Our group is particularly interested in understanding the dynamics and energetics of the switch. While the dynamics are still being interrogated, the energetics of the myristoyl switch for WT and several mutations to residues in the core of the protein — where the myristoyl is thought to bind — have already been determined [240]. The mutation of phenylalanine 6 to leucine (F6L), appears to enhance interactions of myristoyl group within the protein core thereby increasing thermodynamic stability and reducing the ΔG_{switch} compared to WT. Isoleucine 85 to leucine (I85L) results in a “broken” switch where the ΔG_{switch} is ∼0 and changes in pH presumably have no impact on the position of the myristoyl group. Though there is no definitive experimental information concerning exactly how the myristoyl group behaves under these conditions, comparison of changes in thermodynamic stability upon myristoylation suggests it may be “stuck” in the accessible state. I93L, while behaving as WT, can be combined with the previous mutants to generate a triple mutant (F6L/I85L/I93L), which, like I85L alone, has a “broken” switch. In this case, however, the suggestion from experiment is that the myristoyl group is “stuck” in the sequestered state. Notably, a double mutant, I85L/F6L behaves much like WT and it is only with the addition of the WT-like I93L mutant that the particular behaviour of the triple mutant arises. It should be noted that while I85L being stuck in the accessible form and the triple mutant being stuck in the sequestered form are reasonable conclusions from the experimental data [240], it is possible that either is in fact stuck in the opposite state, or in fact, that there is simply no longer an energy difference between states and the myristoyl group freely moves between them without influence of the solution pH. Here, MD is used in an attempt to understand the molecular nature of the myristoyl switch dynamics, specifically, how the aforementioned mutations change the movement of the myristoyl group between states and which interactions are key to promoting those movements.

Initially, equilibrium simulations of: WT, I85L and the triple mutant were performed [240]. These equilibrium simulations agreed with the previously proposed interpretation of the experimental energetics. Specifically, in case of I85L, the myristoyl group preferred to be farther from the bottom of the binding pocket (i.e. more accessible) than WT, while the triple mutant had the opposite preference (Figure 7.6). Relating to the dynamics, I85L showed a much shorter auto-decorrelation time for the distance of the myristoyl group from the bottom of the binding pocket as compared to WT. This suggests faster dynamics of the myristoyl group itself. By contrast, the triple mutant showed much slower dynamics by the same measure (Figure 7.6). Overall, these simulations provided corroboration for the previous experimental conclusions concerning the behaviour of the myristoyl group in the mutant proteins and made suggestions about the impact on dynamics. Still, we sought a more detailed understanding of the dynamics and myristoyl-protein interactions.

Using umbrella sampling, the FES for movement of the myristoyl group within (and out of) the binding pocket was explored, for WT and several mutants (Figure 7.7). While ∼20% of the windows did not reach equilibration as judged by the “sister” method, the total simulation time for those windows reached 60 ns, considerably longer than is typical. Additionally, based on typical techniques of judging equilibration from a single simulation, these would appear to have been at equilibrium. Therefore, I have chosen to construct the FESs from these simulations given that interpretations derived therefrom are no less reliable than may often be seen in the literature. Overall the results show some agreement with experiment and may help to explain some of the experimental findings, but, there is considerable uncertainty in how to compare computational and experimental results. In terms of agreement, the WT FES shows a landscape with two minima, which could be the sequestered and accessible states. In the first, the myristoyl group is fairly deep in the core and the tip makes contacts with F6, I85 and I93 which is in good agreement with results from NMR at high pH where the myristoyl should be in the sequestered state. In the second, while the myristoyl group is still contacting the protein the tip is also partially exposed to solvent and could easily interact with a lipid membrane were it nearby, thus, it appears a reasonable model of the accessible state. Additionally, the FESs show stronger binding of the myristoyl group within the core for F6L and much weaker binding for I85L, which agrees with expectations from experiment. Moreover, I85L shows the smallest energy barrier between the sequestered and accessible states and the sequestered state itself is shifted to being less deeply bound than is the case for WT, all of which agrees with the experimental suggestion that I85L is broken in the accessible state. By contrast, the results for I93L and the triple mutant do not agree as well. I93L, while having a very similar shape to WT, appears to bind less tightly, which should not be the case. The triple mutant is perhaps the most interesting contradiction to experimental expectations, particularly as it could present an alternative explanation of the experimental data. In the case of the triple mutant, the FES indicates a fairly flat energy landscape which would agree with a "broken" switch between sequestered and accessible states. This would suggest the myristoyl group is freely exploring many different binding depths, whereas the existing interpretation of the experimental results suggests it is stuck in the sequestered state [240]. Unfortunately, it is difficult to determine which interpretation is correct without more experimental information. This is because none of the experimentally determined energetics map exactly to values determined from the FES. For instance, the FES can give an estimate of how strong the binding of the myristoyl group is, and what the rates for moving between the sequestered and accessible states would be, but the current experimental results do not give information on the rates or dynamics of the myristoyl switching. Moreover, the experimental energetics resolve the difference in binding between high and low pH, which is not directly comparable to sequestered and accessible at a single pH. Therefore, while considerable resources were devoted to simulating this system, it is unclear if the computational results recapitulate the experimental findings, and as such, it is hard to justify using these results to interpret/understand the molecular behaviour of the system at this point.

The inability to trust these simulations is particularly unfortunate because umbrella sampling effectively produces numerous equilibrium simulations at various points along the reaction coordinate, meaning that it is possible to examine each of these simulations individually in order to gain considerable insight into the behaviour of the system. As an example, the flexibility/dynamics of the protein backbone can be compared as a function of the reaction coordinate. Here I show that when comparing the sequestered to accessible states, there is a clear trend to increased backbone flexibility in the accessible state (Figure 7.8). Given the long simulation times used, there exist a tremendous wealth of such information for WT and the mutants and a myriad of properties could be investigated from the RMSF shown here to changes in rotamer preference, to contact frequencies, hydrogen-bonding, and so forth. Thus, this study can perhaps serve as a warning that considerable forethought ought to go into ensuring a comparison (ideally quantitative) that can be used to be confident that the simulations are recapitulating experiment. It is important to also point out that it is unclear which of the myriad possible variables might provide interesting results. Fortunately, new analysis techniques are providing ways of automatically finding change-points in data of this kind [393].

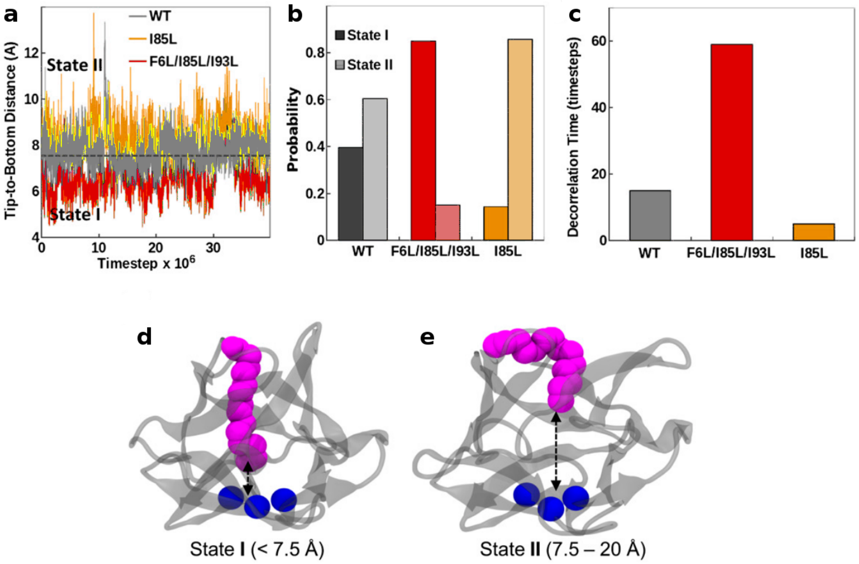


Figure 7.6: Equilibrium simulations of myristoyl dynamics within the binding pocket for WT hisactophilin and mutants. Simple equilibrium simulations were performed to examine the dynamics of the myristoyl group within the binding pocket of hisactophilin. **(a)** The distance of the myristoyl tip to the bottom of the binding pocket is shown across the 80 ns simulation time for WT, I85L and the F6L/I85L/I93L mutants. This distance is the same reaction coordinate used in later umbrella sampling. A black dotted line shows the division between two hypothetical states, I and II. **(b)** The frequency or probability of finding the myristoyl group in each state is shown for WT and each mutant. **(c)** The decorrelation time (of the autocorrelation function) for the distance shown in a, is shown for WT and each mutant. **(d)** Representative structure of "State I" from the published work of Shental-Bechor *et al.* [240], which is roughly equivalent to the "sequestered" state identified in later umbrella sampling. **(e)** Representative structure of "State II" from the published work of Shental-Bechor *et al.* [240], which is roughly equivalent to the "accessible" state identified in later umbrella sampling.

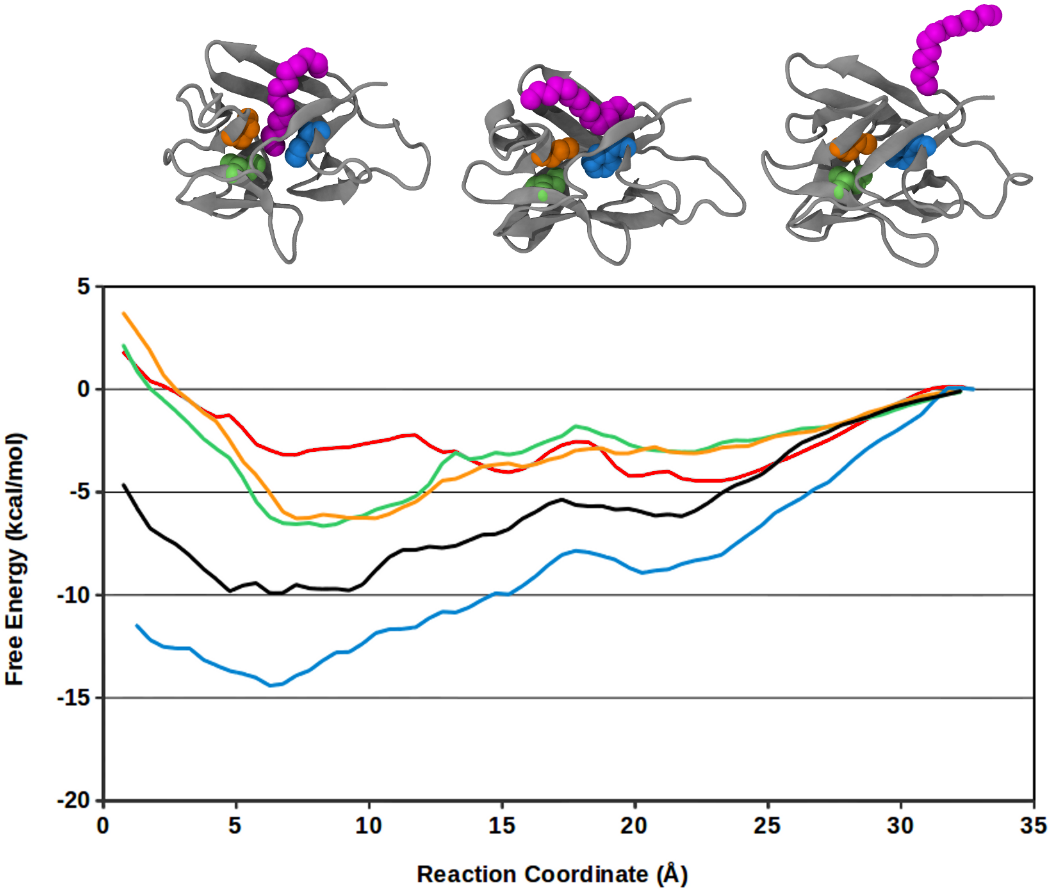


Figure 7.7: FES of myristoylated WT hisactophilin and mutants. FES generated from umbrella sampling are shown for myrisoylated form of the WT (black) protein, and F6L (blue), I85L (orange), I93L (green), and F6L/I85L/I93L (red) mutants. The reaction coordinate is the distance of the tip of the myristoyl group to the bottom of its binding pocket (the hairpin-triplet). Representative structures from MD of the WT protein along the reaction coordinate are shown with WT residues colored in the same manner as the FES plots. There appears to be an energy minimum fairly deep within the pocket (∼5 to 10 Å, first structure on left), termed the "sequestered" state. Additionally there is a local minimum that occurs with very shallow binding (∼20 to 25 Å, middle structure), termed the "accessible" state. Finally, an example of the fully "exposed" state of the myristoyl group is shown at the right. FES are aligned such that this "exposed" state all have an energy of 0 kcal/mol, since mutations within the binding pocket should not affect this state.

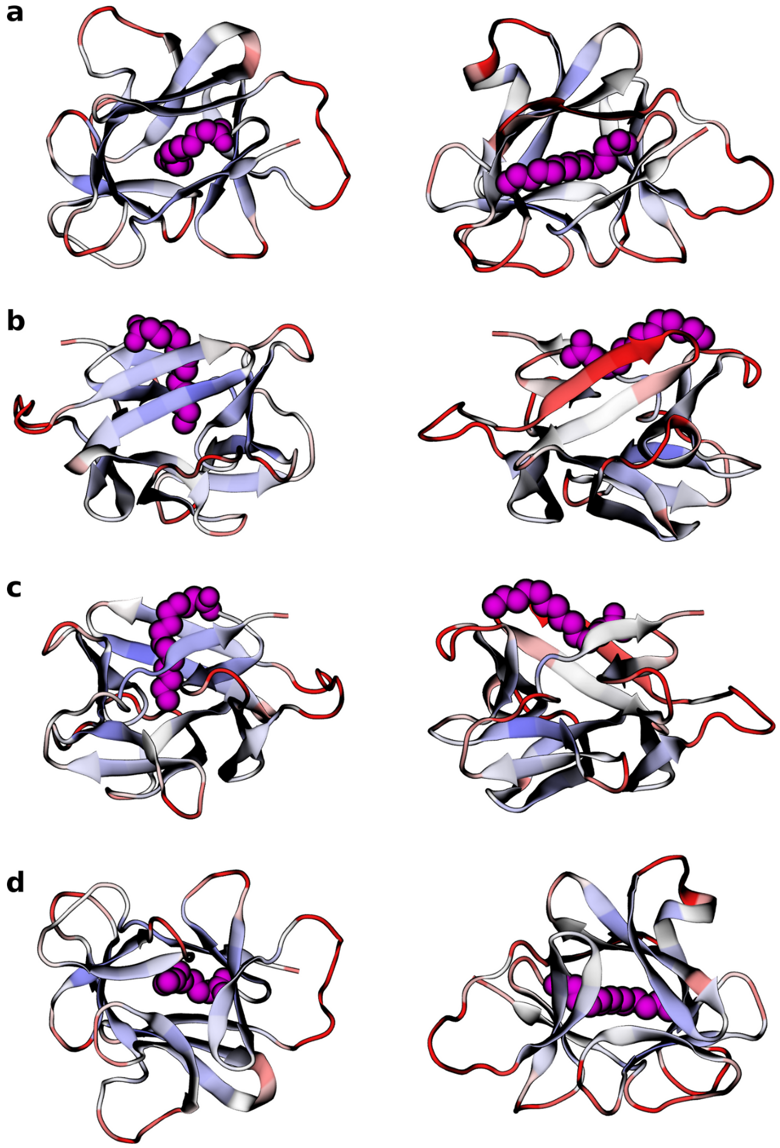


Figure 7.8: Comparison of structural fluctuations between sequestered and accessible forms of WT hisactophilin. The root-mean-squared fluctuations (RMSF) of the backbone atoms for WT myristoylated hisactophilin are shown for both the myristoyl-sequestered (reaction coordinate at ∼9 Å) and myristoyl-accessible (reaction coordinate at ∼22 Å) forms. Blue indicates small fluctuations and red large fluctuations. Views are shown looking down into the barrel **(a)**, from the side closest to the N-terminus **(b)**, from the side closest to the C-terminus **(c)**, and looking down on the hairpin-triplet **(d)**. In general movement of the myristoyl from the sequestered to accessible forms, which is thought to occur at acidic pH, results in nearly universally increased backbone dynamics for beta-structured regions, with particularly high RMSF evident in one of the beta-strands that packs against the accessible myristoyl group.

7.4.7 Metadynamics simulation of a simple water dimer in the gas phase

While the FES for a particular process can be extremely informative, estimation of properties like the binding energy, which depend on the difference in free energy between the bound and fully unbound states can have small errors even for a very well-defined FES. This occurs because the free energy of the "fully unbound" state can only be measured when the ligand is very far from the protein, but typically, the FES is constructed based on simulations where the ligand is only moved far enough from the protein to ablate all direct interactions. That is, where the FES begins to enter a tailing off or plateau. Since the tailing off region is never completely flat, there is always some small additional energy that should be added to the estimate of the binding energy derived from the FES. Simulation of this "long-range" region is far from ideal as little additional information is obtained and thus it is frequently ignored.

Yalina Tritzant-Martinez, working in the lab of Pierre-Nicholas Roy, developed a method by which the FES is fit to a Morse/long-range potential (MLR) in order to provide an exact estimate of the free energy difference when a ligand is fully unbound, as well as allowing the use of transition state theory (TST) to predict on- and off-rates [391]. To test fitting of this potential, a simple ligand system, the gas-phase water dimer, was chosen. The model was initially fit to a "gold standard" (GS) FES, that was obtained by directly iterating over all possible configurations of the water dimer and integrating the energies. Such a direct approach is only possible for a system as simple as the water dimer, thus, a more generally applicable method was also sought to test if the Morse/long-range potential would work as well without such a "gold standard". I assisted in this effort by using metadynamics to construct a FES that might be more representative of typical simulations.

While metadynamics provides a more noisy FES than the GS, as expected and in fact desired, fitting of the MLR potential is little affected, regardless of simulation temperature, or number of parameters used in fitting (Figure 7.9). Determination of rate constants at several temperatures showed excellent agreement between GS and metadynamics (Figure 7.10). Note that this could be used to estimate activation energies, which had been the goal of simulations with the β-lactoglobulin and scFv systems, but the substantial difference in smoothness (and thus accuracy of fitting) of the FESs is strikingly evident (comparing Figure 7.5 to Figure 7.9).

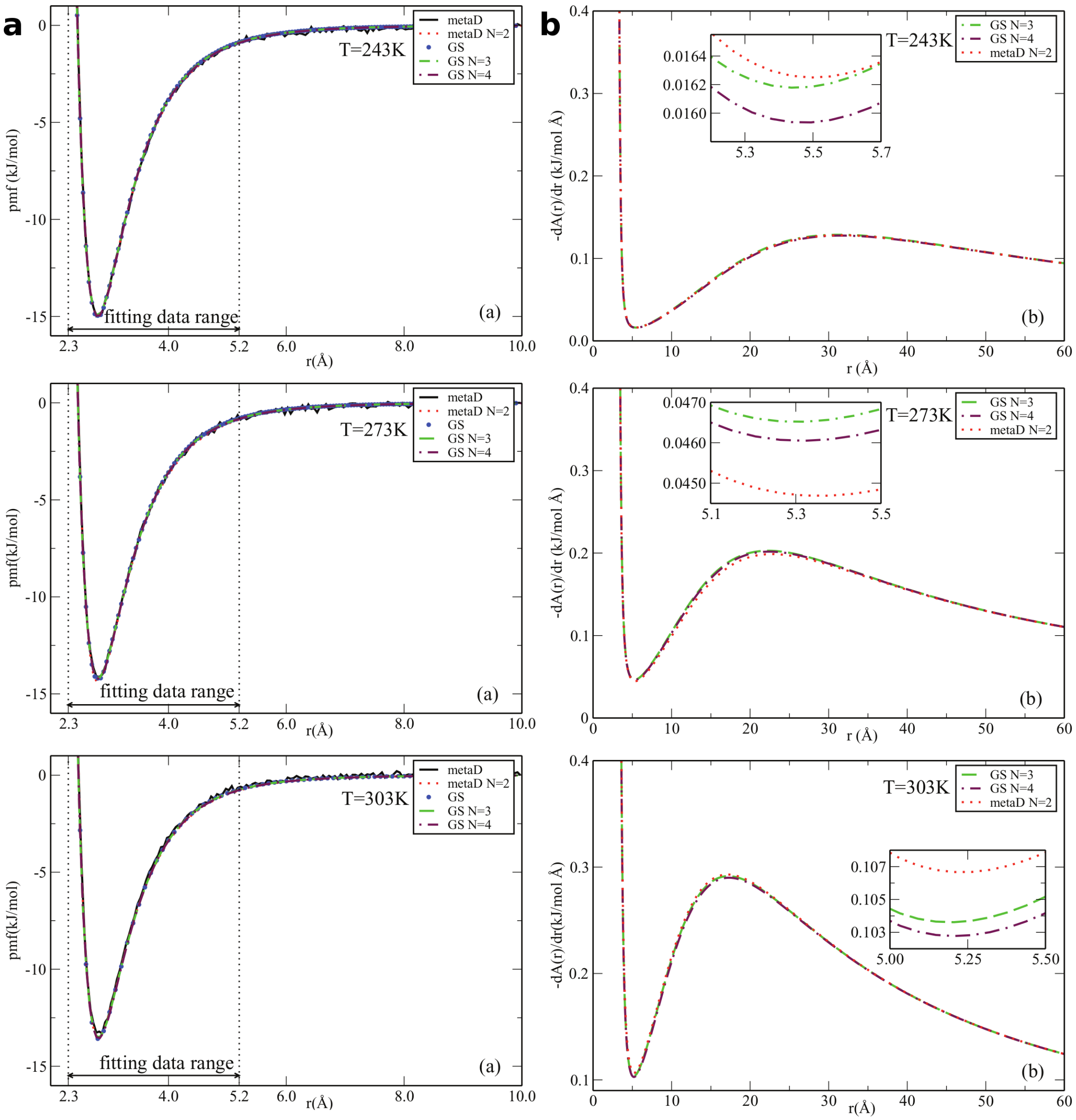


Figure 7.9: Fitting of a Morse/long-range potential to water dimer FES. **(a)** The "gold standard" (GS) values for the water dimer FES are shown as blue circles. The FES determined from metadynamics (metaD) is shown as a blue solid line. Morse/long-range potential (MLR) fits to the metadynamics FES with an N-parameter of 2 (see the work of Tritzant-Martinez *et al.* [391]) is shown with a red dotted line. MLR fits to the GS data with an N-parameter of 2 and 3 are shown with green and purple dotted/dashed lines. Data is shown for three temperatures: 243, 273, and 303 Kelvin. **(b)** The labelling is the same as in (a) but only the -dA(r)/dr (variation of the slope of the Helmholtz free energy along the reaction coordinate) of the MLR fits are represented. The insets represent a magnification near the minima of the curves from 5.0 to 5.5 Å. Figure taken from [391].

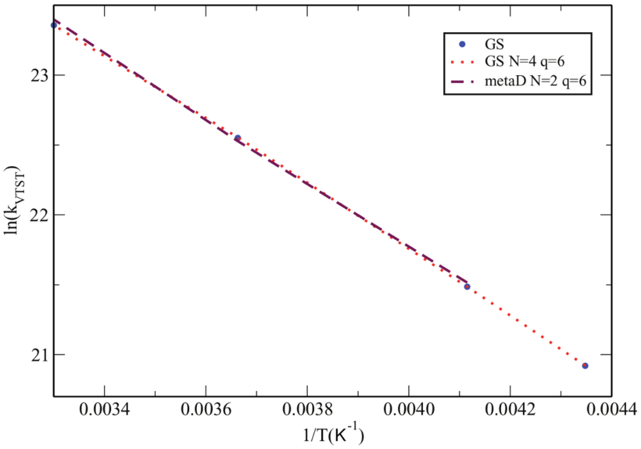


Figure 7.10: Temperature dependence of rate constant from MLR fitted models. Values of $\ln(k_{VTST})$ versus $1/T$ calculated for the GS, an MLR fitted model of GS with N = 4, q = 6, and an MLR fitted model of the metadynamics calculation with N = 2 and q = 6.

7.5 Discussion

MD, and biased MD in particular, are powerful tools, but considerable care needs to be taken if they are to be used effectively. Questions of equilibration/convergence are paramount and generally mean that interrogating very complex interactions with many possible degrees of freedom is a fraught endeavour not to be approached lightly. Nevertheless, if such a task is to be attempted, methods like the "sister" simulation outlined here, provide a means of identifying cases where equilibration is not occurring.

Here, three protein-ligand systems were examined in the gas phase (no solvent) using both simple equilibrium simulations (so called "vanilla molecular dynamics") and biased MD. For the streptavidin-biotin system in the gas-phase the simulations showed that the interactions present in the gas-phase are nearly identical to those in solution. Based on experiment, it is likely this results from the high kinetic stability (high activation free energy for dissociation) of the protein-ligand complex in the gas-phase. Umbrella sampling of β-lactoglobulin with its fatty acid ligands, which was examined by Nobar Jalili — using scripts I had written for the automation of many aspects of the umbrella sampling protocol — gave results that did not agree with experiment, and Nobar's conclusion was that long-lived (kinetically stable) electrostatic interactions between the fatty acid headgroup and charged residues on the protein surface caused poor exploration and equilibration of the protein-ligand interactions along the binding reaction coordinate [389]. In my own umbrella sampling simulations of the scFv antibody fragment (and binding residue mutant H228A) with its trisaccharide ligand (and modified deoxy-mannose ligand), I found results with excellent qualitative agreement to experiment. Unfortunately, the noise in the associated FES and the long simulation times (and thus high computational cost) needed to get those results meant that a complete set of *in silico* experiments on this system was not possible. It is likely that the long simulation times needed for equilibration of many of the umbrella sampling windows resulted from kinetically stable electrostatic interactions much like what Nobar identified. Since the trisaccharide ligand does not have fully charged atoms, but only partially charged polar groups, the effect of this kinetic stability is likely reduced. Therefore, while the kinetic stability of protein-ligand interactions in the gas phase has been a boon for *in vitro* experiments, allowing the use of mass spectrometry to determine binding constants and identify interactions that were present in solution, this same kinetic stability is a bane for *in silico* experiments were it makes the equilibration and convergence needed to get reliable results extremely challenging.

A single protein-ligand system (or pseudo-ligand), the myristoyl-hisactophilin system, was examined using explicit solvent simulations and umbrella sampling. These simulations, while very computationally costly owing to the simulation of thousands of water molecules, gave results with moderate agreement to experiment. Interaction of the myristoyl group with the protein core is largely composed of non-polar interactions and is not expected to suffer from the same problems as seen with the gas-phase simulations. Nonetheless, equilibration was slow, and some windows did not equilibrate, which certainly contributes to the discrepancies between simulation and experiment (though other factors like forcefield inaccuracies could also play a role). In this case, while I cannot be certain why the simulations are slow to equilibrate, the highly flexible nature of the myristoyl group and its interactions with many flexible sidechains within the hydrophobic core of the protein suggest that the sheer number of degrees of freedom involved in binding might be the primary problem. Looking back at the gas-phase simulations would suggest that the much poorer results seen with the β-lactoglobulin and fatty acid system than the scFv trisaccharide system, might be due in part to the similarly long and flexible fatty acid ligand binding a deep pocket with many interactions. Thus, examination of protein-ligand systems with large flexible ligands and deep pockets is likely to be much more problematic than smaller or more rigid ligands interacting with shallow grooves. In the aforementioned streptavidin-biotin system, while streptavidin has a relatively deep binding pocket, a substantial portion of the biotin molecule is fairly rigid, being composed of a double ring system. While not shown in this thesis (though in preparation for publication), I did perform umbrella sampling simulations on the phosphoenolpyruvate carboxykinase system, examining the FES for enzymatic lid opening and closing with different mutations. Such simulations appeared to agree well with experiment and equilibration using the “sister” simulation method developed here occurred for all windows. Therefore, if any unifying theme is to emerge from all of these umbrella sampling simulations it is that gas-phase simulations — while alluring owing to the low computational cost of avoiding solvent — are best avoided due to potentially high kinetic barriers, and systems involving surface interactions (like the scFv trisaccharide, or an enzymatic lid) are preferable over those with interactions deep within the protein (as seen with β-lactoglobulin and hisactophilin with their fatty-acid ligands.

While metadynamics on the water-dimer was successful, and certainly much simpler to setup and execute than umbrella sampling would have been, this is an extraordinarily simple system, and the simulation of proteins is still plagued with equilibration and convergence problems even with metadynamics [383] and other similar approaches [384]. Still, the concepts behind these approaches may help illuminate the behaviour of systems which are not not accessible without the use of an artificial bias, and require reaction coordinates with multiple dimensions.

7.6 Methods

7.6.1 Simulations of the streptavidin-biotin complex

MD simulations were performed using the AMBER 11 program suite (Accelrys, San Diego, CA). As there is no available crystal structure of the WT streptavidin-biotin tetramer, the initial geometry of the WT (S_4+4B) complex was obtained by applying a crystallographic symmetry operator on the crystal structure of WT streptavidin-biotin dimer (PDB ID: 3RY2). Each of the tetrameric chains (containing residues 14-134 for the A and C chains, residues 15-136 for the B and D chains) was extended to have the same length as the truncated form (containing residues 13-139) of the WT streptavidin used experimentally. This was done by aligning each tetrameric chain against the C-chain of the crystal structure for streptavidin mutant Ser27Ala (PDB ID: 1N9M) using the DALI server [394] and grafting the extended residues onto each chain of the tetramer. Currently with AMBER, atomic charges and atom type parameters are available only for the charged forms of the Arg, C-terminal Ser, and N-terminal Ala residues. Consequently, it was necessary to develop charges and parameters for the neutral forms of Arg, C-terminal Ser, and N-terminal Ala. The charge parameters of the neutral forms of Arg, C-terminal Ser, and N-terminal Ala were parametrized as tripeptides (NME-Arg-ACE, NME-Gly-CSer, and NAla-Gly-ACE, respectively) using the RESP ESP charge derive server (RED Server) [395, 396] using Gaussian C.01 and enforcing net neutrality across the residue. Ions of the WT (S_4+4B) complex at the +12 charge state were chosen for investigation. As described in more detail below, 15 different charge distributions were considered.

Topology and coordinate files for the simulations of each charge distributions were created using the AntechAMBER module of the AMBERTools (version 11) [397]. The MD simulations were performed using the AMBER 03 force field for streptavidin and a general AMBER force field (GAFF) [398] for Biotin. Simulations were performed using a 2 fs time step with bonds to hydrogens constrained using SHAKE [369]. The NVT ensemble was used with an Anderson thermostat (300 K, collision frequency 1 ps^{-1}) and no nonbonded cutoff (full nonbonded interactions). The system was minimized using 500 steps of steepest descent and 500 steps of conjugate gradient minimization. Following minimization, 10 ns of dynamics was needed to fully equilibrate the system, as judged by Cα root-mean-square deviation (RMSD). Following this 10 ns of equilibration, a further 10 ns of production was performed and used in the analysis.

7.6.2 Analysis of streptavidin-biotin simulations

Trajectory analysis, performed using the Visual Molecular Dynamics package (http://www.ks.uiuc.edu/Research/vmd/), was carried out to establish the Cα RMSD for streptavidin, the angles and distances associated with the H-bonds between biotin and streptavidin, and the distances associated with the intermolecular van der Waals interactions between biotin and four tryptophan residues, Trp 79, Trp92, Trp108, and (adjacent subunit) Trp120. In order to determine the number of H-bonds, all potential H-bonding partners between biotin and streptavidin were scanned at each frame, with the criteria for an H-bond being a heavy-atom distance $\leq$4 Å and a donor-hydrogen-acceptor angle $\leq$120°. The total number of H-bonds for each configuration was then averaged across all frames and the four subunits. Additionally, the occupancy, i.e., the fraction of the simulation steps for which the H-bond criteria are satisfied, was evaluated. All potential van der Waals interaction atom pairs between biotin and four tryptophan residues (Trp 79, Trp92, Trp108, and Trp120 from adjacent subunit) were scanned at each frame, with the criterion for the presence of van der Waals interactions being that the distance between atom centers for each pair of atoms between relevant tryptophan residue and biotin is less than or equal to the sum of the van der Waals radii for those particular atoms (van der Waals radii based on the AMBER parameter set were used for the simulations). The fraction of the simulation steps for which the van der Waals interaction criteria are satisfied was also evaluated.

7.6.3 Fatty acid and non-standard residue parameterization

While the AMBER12SB and complementary GLYCAM06 [399] forcefields provide all parameters needed for simulating proteins and carbohydrates under solution conditions, the gas-phase simulations required some non-standard residues as well as parameters for the fatty acids used with β-lactoglobulin. In particular, non-standard residues were needed because AMBER12SB does not include parameters for neutral arginine, nor for any neutral C- or N-terminal amino acids. For both fatty acids and these non-standard residues, the partial charges on each atom were computed using the RESP ESP charge derive server with Guassian 09 (C.01). Geometry optimization was accomplished using the RESP (Restrained Electrostatic Potential) methods at an HF/6-31G* level of theory. For fatty acids, the partial charges were then used along with the General AMBER Forcefield (GAFF) [398] to produce MD parameters. For the neutral C- and N-terminal residues the partial charges were used to modify the existing standard terminal residue types within the AMBER12SB forcefield.

7.6.4 Choice of charged residues for gas-phase simulations

Two methods were developed for determining an optimum charge configuration that would reflect that of the protein ionized for mass spectrometry. In both cases the protein structure is read from a PDB file reflecting the solution structure of the protein and the final goal is to minimize the electrostatic potential energy.

The first method, used for the β-lactoglobulin system, involved using AMBER's LeAP program to add hydrogens to the coordinates based on the AMBER12SB forcefield, followed by evaluation of the electrostatic potential energy using custom code. The all-atom structure output by LeAP was loaded into a custom program written in C++ which loads the AMBER12SB forcefield and then converts all: Asp, Glu, His, Lys, Arg, and C- and N-terminal groups to their uncharged/neutral form, thus rendering the entire protein neutral. If the net charge during mass spec is negative (as is the case for β-lactoglobulin) then negative ionization is only considered for: Asp, Glu and the C-terminus. If the net charge is positive, then positive ionization is only considered for: His, Lys, Arg, and the N-terminus. Given the total net charge from experiment, the program then constructs a list of all possible permutations for the location of those charges (for β-lactoglobulin this was 888,030 possibilities). Each permutation then had its electrostatic potential energy fully evaluated. Custom code was necessary in this case because even though any MD program could evaluate the electrostatic potential energy quickly, the construction of 888,030 individual input files would have taken up an extremely large amount of space and considerable computational time to generate in the first place. In general MD programs are not intended for this kind of analysis.

The best 1% (8880) charge configurations (lowest electrostatic potential energy) from the previous step were chosen for more detailed analysis. Using the AMBER12SB forcefield and the NAMD program, conjugate-gradient energy minimization was performed for 5000 steps on each configuration. From this the configuration with the lowest combined electrostatic and van der Waals potential energy was chosen for the main umbrella sampling analysis. Energy minimization was used to allow for rotation of sidechains to more optimal positions given that the charge configurations being tested were different than that in the original crystal structure. The consideration of van der Waals forces was added at the final stage to ensure that energy minimization did not produce a configuration that appeared ideal from an electrostatic point of view, but would actually have been unfavourable overall.

The second method, used for the scFv system, was intended as a faster and more easily usable approach. In this case, the structure (again loaded from the PDB file) is simplified to just the ionizable atoms of each sidechain (e.g. the terminal nitrogen of Lysine in the case of a positive net-charge). All possible permutations are still considered, but the evaluation of electrostatic potential energy is much faster as only a fraction of the calculations are needed. While this is certainly less accurate, it nevertheless produces configurations where the charged groups are well-distributed about the structure. Given that the experimental conditions likely generate an ensemble of different charge configurations anyway, this method likely produces a roughly comparable solution to the more detailed one above.

7.6.5 Simulation conditions for β-lactoglobulin and scFv

Simulations of both the β-lactoglobulin and scFv systems were performed in a vacuum, that is: no solvent, no barostat, and no periodic boundary conditions. Because of the lack of periodic boundary conditions, full electrostatics were calculated (no cutoff). A timestep of 2 fs was used with the SHAKE algorithm constraining bonds to hydrogen. For the β-lactoglobulin system a temperature of 299 K was maintained (to mimic experimental temperatures) using Langevin-dynamics with a collision frequency of 1 ps^{-1}. For the scFv system a temperature of 390 K was used with the same Langevin-dynamics parameter. For both systems the AMBER12SB forcefield was used and for the trisaccharide the GLYCAM06 forcefield. For β-lactoglobulin NAMD was used, whereas for scFv, code to automate umbrella sampling was written using OpenMM (https://simtk.org/projects/openmm) [400].

7.6.6 Simulation conditions for hisactophilin

Because the hisactophilin simulation involves a covalently modified N-terminus (myristoylation), partial charges for an N-terminally linked myristoyl group were derived as noted in section 7.6.3. For the protein, AMBER03 was used, and for parameterizing the myristoyl group GAFF was used as in the fatty acids in section 7.6.3. The linkage between the myristoyl group and the remainder of the protein was given atom types based on the AMBER03 forcefield.

For equilibrium simulations, which were used in the work of Shental-Bechor *et al.* [240] (Figure 7.6), were performed using the AMBER simulation package (http://ambermd.org/). The TIP3P water model was used for explicit solvation of 8 Å around the protein. All hydrogens were constrained to have rigid bonds. A long-range cutoff of 10 Å was used with periodic boundary conditions and long-range electrostatics modelled using the Particle mesh Ewald approximation. The initial model of the myristoylated protein was built from the NMR structure (PDB 1HCD) with the myristoyl group manually modelled based on NMR restraints. Coordinates were initially minimized by conjugate-gradient energy minimization. Simulations were performed using Langevin dynamics at 298 K with a 2 fs timestep and a Berdnensen barostat at 1 atm. The simulations were first equilibrated for 20 ns and followed by 80 ns of production.

For umbrella sampling simulations, the simulation conditions were the same as above, but the NAMD package was used for simulations. Initial equilibration was performed for 100 ns, and starting coordinates at 120 and 140 ns were used for subsequent umbrella sampling simulations. As noted in the results, umbrella sampling simulations proceeded in segments of 10 ns.

7.6.7 Particulars of umbrella sampling

The umbrella sampling reaction coordinate in the case of the β-lactoglobulin system was the distance between the center of mass of the fatty acid headgroup and the center of mass of the Cα atoms of the barrel that forms the binding pocket. For the scFv system the reaction coordinate was the distance between the center of mass of Abe sugar residue and the center of mass of the Cα atoms of the middle beta-barrel of the protein (which the ligands essentially binds on top of). Note that a conical restraint was used in the case of the scFv system to only allow the ligand to exit the binding site within a cone with a 60° angle. This was done to reduce non-specific binding of the ligand to the protein during umbrella sampling. For the hisactophilin system the reaction coordinate was the distance between center of mass of the last two carbon atoms of the myristoyl group and the center of mass of the Cα atoms from valines 21, 61, and 101, which are located symmetrically at the bottom of the myristoyl binding pocket.

Umbrella sampling windows for the β-lactoglobulin system were spaced 0.25 Å apart over a distance of 30 Å and then 0.5 Å apart over another 30 Å for a total of 180 windows. For the scFv system, windows were spaced 0.5 Å apart over a distance of 34 Å for a total of 68 windows. For the hisactophilin system windows were spaced 0.5 Å apart over a distance of 31.5 Å for a total of 63 windows. Since each window was simulated twice, and the minimum simulation time was 10 ns, the minimum total simulation time for the β-lactoglobulin, scFv, and hisactophilin systems was 3600 ns, 1360 ns, and 1260 ns respectively. Since many windows took much longer than the base amount of time to equilibrate, the actual total times are at least twice as long in all cases.

7.6.8 Metadynamics

Metadynamics of the water dimer system were performed in the gas phase using the TIP4P water model, and bonds to hydrogen made rigid. Full electrostatics were calculated (no long-range cutoff), and an integration timestep of 2 fs was used with Langevin dynamics and a 1 ps^{-1} collisional frequency. The height of the repulsive hills being added was 0.001 kcal/mol based on a reaction coordinate between the center of mass of each water molecule. New hills were added every 500 simulation steps. Simulations were run for 100 ns in total.

Chapter 8

Discussion and Future Work

8.1 Computational protein design: what was learned and where to go next?

Using existing sequence and structure databases to inform protein design has been very successful as seen in the design of ThreeFoil (see Chapter 2), and other designs [45, 47, 67, 54]. Similarly, the use of modularity and symmetry to simplify the design likely contributed to the successful design of ThreeFoil and other symmetric or repeated proteins [127, 45, 46, 47, 21, 19, 22, 54]. Interestingly, a single repeat of ThreeFoil (OneFoil) did not fold or self-assemble [44], whereas one third of the Pizza6 protein was able to fold [45]. Part of this may be that the β-trefoil fold, in being globular, is a more complex modular assembly target than that of the toroidal β-propeller fold. Interestingly, however, the Pizza6 protein was initially designed in Rosetta as a homo-hexamer and only later optimized as a monomer [45], whereas ThreeFoil was designed as a single symmetric monomer from the outset. Thus, one naturally wonders if the design lineage for the β-propeller case was analogous to the evolutionary lineage, and this allowed better recapitulation of evolutionary intermediate structures. Analysis of the β-prism fold suggests it has similarly arisen from symmetric expansion (see Chapter 3). Therefore, it is interesting to consider to possibility of an effective design process based on that for ThreeFoil or Pizza, but taking into account everything that has been learned since (see Chapter 1) in order to design a completely symmetric β-prism.

Having examined the topological complexity of a number of folds (see Chapters 4 and 5), it is notable how often elements of secondary structure appear to be ordered in space so as to maximize this complexity. In fact, it has previously been noted that the preponderance of folds in which the N- and C-termini interact is substantially higher than expected by chance, and that such an arrangement is particularly evident for two-state folders [401]. At the time it was suggested that this may play a role in folding, stability and turnover. The results presented in Chapters 4 and 5 agree with these conclusions, that is, such an "N-terminal to C-terminal motif" likely improves unfolding cooperativity and kinetic stability not only improving protein half-life but reducing the opportunity for wanton aggregation. While the overall arrangement of strands in the β-trefoil fold is fairly complex and ThreeFoil's resistance to harsh conditions originates from this, the topological complexity of the β-prism fold is yet higher still. The individual repeating structural motif of the β-prism is the four-stranded greek-key, which is itself a topologically complex motif, having direct contact between the first and fourth strands. A circular permutation then results in half of the first greek key being N-terminal while the other half is C-terminal further increasing topological complexity. Given the potential value of designing a symmetric β-prism, it would be intriguing to design a variant without this permutation and observe how the resulting change in topological complexity manifests.

The results from Chapter 6, while placed in terms of understanding point mutations and their impact, are nonetheless relevant to protein design. That is, being able to accurately determine the impacts of a residue substitution is essentially the fundamental step of protein design, which is then repeated until an optimum (or at least seemingly optimum) sequence is obtained. To date, Rosetta has been the overwhelmingly dominant (both in terms of usage and success) protein design forcefield (see Chapter 1). The analysis in Chapter 6, however, suggests Rosetta — based on a mixture of physical and statistical potential terms — may be only slightly above average at choosing an optimum residue. Notably, the purely statistical potentials DFire and PoPMuSiC perform better almost universally and are at least as fast. While, LIE and CC/PBSA — which use purely physical potentials — do not universally outperform Rosetta, their Spearman correlation coefficients are higher, indicating they are better at ranking residue choices in terms of stability (even if they do worse at identifying the exact value). It is possible that Rosetta has been particularly successful owing not just to a competent forcefield, but its combination with advanced algorithms accounting for backbone flexibility [402, 403] and improving upon specific common problems [87]. As such, it might be challenging for other forcefields to compete in the protein design game. Chitsaz and Mayo [404], however, may have helped level the playing field with their GRID algorithm with can be hooked up to any all-atom forcefield in order to optimize both sidechain rotamers and backbone dihedrals. Therefore, following on the previous ideas of designing a symmetric β-prism, it would be intriguing and perhaps revealing to compare design using Rosetta to that using DFire, PoPMuSiC, and an MD forcefield-based approach.

In addition to the beneficial design characteristics outlined above, the β-prism fold has its three symmetric binding sites particularly close in space, resulting in three flexible binding loops also close in space, much like an immunoglobulin fold (see Chapter 3). Thus, the β-prism is not only an interesting design case for testing evolutionary hypotheses, design approaches and forcefields, and the role of topological complexity in design, but may also be a valuable binding scaffold in its own right.

8.2 Molecular dynamics of ligand binding: what was learned and where to go next?

Umbrella sampling simulations, which were used to interrogate protein-ligand binding in several systems (see Chapter 7), demonstrate that while the method is very powerful, accurate results may require considerable simulation time in order to reach equilibrium. Therefore, it is critical to approach umbrella sampling, or other biased techniques like metadynamics, with much scepticism when simulation times are short. In particular, when the predicted free energy changes for a process are considerably larger than might be reasonably expected, it is likely that poor equilibration is the problem. Unfortunately, determining when a system has fully equilibrated can be difficult. While I've presented the idea that comparing "sister" simulations that have begun from slightly different starting coordinates can be a useful technique, the associated doubling of simulation time is hardly ideal. Still, umbrella sampling under such stringent conditions reveals itself as a useful, if costly, technique for gaining atomic level insight into the free energy surface of particular protein behaviours. For those interested, my subjective opinion is that systems with complexity on the order of a protein-ligand system, will likely require several thousands of nanoseconds worth of simulation time to get usable results.

References

[1] Broom, A., Trainor, K., MacKenzie, D. W., and Meiering, E. M. (2016) Using natural sequences and modularity to design common and novel protein topologies. *Curr Opin Struct Biol*, **38**, 26–36.

[2] Alberts, B., Johnson, A., Lewis, J., David, M., Raff, M., Roberts, K., and Walter, P. (2014) *Molecular Biology of the Cell*. Garland Science, 6 edn.

[3] Khan, S. and Vihinen, M. (2007) Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct Biol*, **7**, 56.

[4] Alcalde, M., Ferrer, M., Plou, F. J., and Ballesteros, A. (2006) Environmental biocatalysis: from remediation with enzymes to novel green processes. *Trends Biotechnol*, **24**, 281–287.

[5] Trudeau, D. L., Lee, T. M., and Arnold, F. H. (2014) Engineered thermostable fungal cellulases exhibit efficient synergistic cellulose hydrolysis at elevated temperatures. *Biotechnol Bioeng*, **111**, 2390–2397.

[6] Orozco, J., Vilela, D., Valdés-Ramírez, G., Fedorak, Y., Escarpa, A., Vazquez-Duhalt, R., and Wang, J. (2014) Efficient biocatalytic degradation of pollutants by enzyme-releasing self-propelled motors. *Chemistry*, **20**, 2866–2871.

[7] Ran, N., Zhao, L., Chen, Z., and Tao, J. (2008) Recent applications of biocatalysis in developing green chemistry for chemical synthesis at the industrial scale. *Green Chemistry*, **10**, 361–372.

[8] Lipovsek, D. (2011) Adnectins: engineered target-binding protein therapeutics. *Protein Eng Des Sel*, **24**, 3–9.

[9] Tamura, T. and Hamachi, I. (2014) Recent progress in design of protein-based fluorescent biosensors and their cellular applications. *ACS Chem Biol*, **9**, 2708–2717.

[10] Clark, D. and Mao, L. (2012) Cancer biomarker discovery: lectin-based strategies targeting glycoproteins. *Dis Markers*, **33**, 1–10.

[11] yi Shen, M., Davis, F. P., and Sali, A. (2005) The optimal size of a globular protein domain: A simple sphere-packing model. *Chemical Physics Letters*, **405**, 224–228.

[12] Kolodny, R., Pereyaslavets, L., Samson, A. O., and Levitt, M. (2013) On the universe of protein folds. *Annu Rev Biophys*, **42**, 559–582.

[13] Bryan, P. N. and Orban, J. (2010) Proteins that switch folds. *Curr Opin Struct Biol*, **20**, 482–488.

[14] Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N., and Mayo, S. L. (2012) Iterative approach to computational enzyme design. *Proc Natl Acad Sci U S A*, **109**, 3790–3795.

[15] Li, Z., Yang, Y., Zhan, J., Dai, L., and Zhou, Y. (2013) Energy functions in de novo protein design: current challenges and future prospects. *Annu Rev Biophys*, **42**, 315–335.

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

[17] Dantas, G., Kuhlman, B., Callender, D., Wong, M., and 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.

[18] Watters, A. L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., and Baker, D. (2007) The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell*, **128**, 613–624.

[19] Thomson, A. R., Wood, C. W., Burton, A. J., Bartlett, G. J., Sessions, R. B., Brady, R. L., and Woolfson, D. N. (2014) Computational design of water-soluble alpha-helical barrels. *Science*, **346**, 485–488.

[20] Taylor, W. R. (2002) A 'periodic table' for protein structures. *Nature*, **416**, 657–660.

[21] Huang, P.-S., Oberdorfer, G., Xu, C., Pei, X. Y., Nannenga, B. L., Rogers, J. M., DiMaio, F., Gonen, T., Luisi, B., and Baker, D. (2014) High thermodynamic stability of parametrically designed helical bundles. *Science*, **346**, 481–485.

[22] Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. (2015) Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol*, **22**, 167–174.

[23] Brunette, T. J., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. (2015) Exploring the repeat protein universe through computational protein design. *Nature*, **528**, 580–584.

[24] Woolfson, D. N., Bartlett, G. J., Burton, A. J., Heal, J. W., Niitsu, A., Thomson, A. R., and Wood, C. W. (2015) De novo protein design: how do we expand into the universe of possible protein structures? *Curr Opin Struct Biol*, **33**, 16–26.

[25] Currin, A., Swainston, N., Day, P. J., and Kell, D. B. (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem Soc Rev*, **44**, 1172–1239.

[26] Jacobs, S. A., Diem, M. D., Luo, J., Teplyakov, A., Obmolova, G., Malia, T., Gilliland, G. L., and O'Neil, K. T. (2012) Design of novel fn3 domains with high stability by a consensus sequence approach. *Protein Eng Des Sel*, **25**, 107–117.

[27] Porebski, B. T., Nickson, A. A., Hoke, D. E., Hunter, M. R., Zhu, L., McGowan, S., Webb, G. I., and Buckle, A. M. (2015) Structural and dynamic properties that govern the stability of an engineered fibronectin type iii domain. *Protein Eng Des Sel*, **28**, 67–78.

[28] Aerts, D., Verhaeghe, T., Joosten, H.-J., Vriend, G., Soetaert, W., and Desmet, T. (2013) Consensus engineering of sucrose phosphorylase: the outcome reflects the sequence input. *Biotechnol Bioeng*, **110**, 2563–2572.

[29] Risso, V. A., Gavira, J. A., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2014) Phenotypic comparisons of consensus variants versus laboratory resurrections of precambrian proteins. *Proteins*, **82**, 887–896.

[30] Wheeler, L. C., Lim, S. A., Marqusee, S., and Harms, M. J. (2016) The thermostability and specificity of ancient proteins. *Curr Opin Struct Biol*, **38**, 37–43.

[31] Lees, J., Dawson, N., Sillitoe, I., and Orengo, C. (2016) Functional innovation from changes in protein domains and their combinations. *Curr Opin Struct Biol*, **38**.

[32] Tóth-Petróczy, A. and Tawfik, D. S. (2014) The robustness and innovability of protein folds. *Curr Opin Struct Biol*, **26**, 131–138.

[33] Rockah-Shmuel, L., Ágnes Tóth-Petróczy, and Tawfik, D. S. (2015) Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput Biol*, **11**, e1004421.

[34] Magliery, T. J. (2015) Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol*, **33**, 161–168.

[35] Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774–786.

[36] Marks, D. S., Hopf, T. A., and Sander, C. (2012) Protein structure prediction from sequence variation. *Nat Biotechnol*, **30**, 1072–1080.

[37] Sullivan, B. J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., Syu, T., and Magliery, T. J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J Mol Biol*, **420**, 384–399.

[38] Murphy, G. S., Mills, J. L., Miley, M. J., Machius, M., Szyperski, T., and Kuhlman, B. (2012) Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure*, **20**, 1086–1096.

[39] Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature*, **491**, 222–227.

[40] Söding, J. and Lupas, A. N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, **25**, 837–846.

[41] Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2014) Global view of the protein universe. *Proc Natl Acad Sci U S A*, **111**, 11691–11696.

[42] Pitman, D. J., Schenkelberg, C. D., Huang, Y.-M., Teets, F. D., Ditursi, D., and Bystroff, C. (2014) Improving computational efficiency and tractability of protein design using a piecemeal approach. a strategy for parallel and distributed protein design. *Bioinformatics*, **30**, 1138–1145.

[43] Balaji, S. (2015) Internal symmetry in protein structures: prevalence, functional relevance and evolution. *Curr Opin Struct Biol*, **32**, 156–166.

[44] Broom, A., Doxey, A. C., Lobsanov, Y. D., Berthin, L. G., Rose, D. R., Howell, P. L., McConkey, B. J., and Meiering, E. M. (2012) Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure*, **20**, 161–171.

[45] Voet, A. R. D., Noguchi, H., Addy, C., Simoncini, D., Terada, D., Unzai, S., Park, S.-Y., Zhang, K. Y. J., and Tame, J. R. H. (2014) Computational design of a self-assembling symmetrical beta-propeller protein. *Proc Natl Acad Sci U S A*, **111**, 15102–15107.

[46] Huang, P.-S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Höcker, B., and Baker, D. (2016) De novo design of a four-fold symmetric tim-barrel protein with atomic-level accuracy. *Nat Chem Biol*, **12**, 29–34.

[47] Parmeggiani, F., et al. (2015) A general computational approach for repeat protein design. *J Mol Biol*, **427**, 563–575.

[48] Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.

[49] Lobb, B. and Doxey, A. C. (2016) Novel function discovery through sequence and structural data mining. *Curr Opin Struct Biol*, **38**, 53–61.

[50] Höcker, B. (2014) Design of proteins from smaller fragments-learning from evolution. *Curr Opin Struct Biol*, **27**, 56–62.

[51] Bharat, T. A. M., Eisenbeis, S., Zeth, K., and Höcker, B. (2008) A beta alpha-barrel built by the combination of fragments from different folds. *Proc Natl Acad Sci U S A*, **105**, 9942–9947.

[52] Höcker, B., Claren, J., and Sterner, R. (2004) Mimicking enzyme evolution by generating new (betaalpha)8-barrels from (betaalpha)4-half-barrels. *Proc Natl Acad Sci U S A*, **101**, 16448–16453.

[53] Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R., and Sterner, R. (2010) Computational and experimental evidence for the evolution of a (betaalpha)8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds. *J Mol Biol*, **398**, 763–773.

[54] Smock, R. G., Yadid, I., Dym, O., Clarke, J., and Tawfik, D. S. (2016) De novo evolutionary emergence of a symmetrical protein is shaped by folding constraints. *Cell*, **164**, 476–486.

[55] Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat Struct Biol*, **9**, 553–558.

[56] Trudeau, D. L., Smith, M. A., and Arnold, F. H. (2013) Innovation by homologous recombination. *Curr Opin Chem Biol*, **17**, 902–909.

[57] Kufner, K. and Lipps, G. (2013) Construction of a chimeric thermoacidophilic beta-endoglucanase. *BMC Biochem*, **14**, 11.

[58] van Beek, H. L., de Gonzalo, G., and Fraaije, M. W. (2012) Blending baeyer-villiger monooxygenases: using a robust bvmo as a scaffold for creating chimeric enzymes with novel catalytic properties. *Chem Commun (Camb)*, **48**, 3288–3290.

[59] Zhou, X., Gao, L., Yang, G., Liu, D., Bai, A., Li, B., Deng, Z., and Feng, Y. (2015) Design of hyperthermophilic lipase chimeras by key motif-directed recombination. *Chembiochem*, **16**, 455–462.

[60] Gobeil, S. M. C., Clouthier, C. M., Park, J., Gagné, D., Berghuis, A. M., Doucet, N., and Pelletier, J. N. (2014) Maintenance of native-like protein dynamics may not be required for engineering functional proteins. *Chem Biol*, **21**, 1330–1340.

[61] Rogers, R. L. and Hartl, D. L. (2012) Chimeric genes as a source of rapid evolution in drosophila melanogaster. *Mol Biol Evol*, **29**, 517–529.

[62] Cui, Y., Wong, W. H., Bornberg-Bauer, E., and Chan, H. S. (2002) Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci U S A*, **99**, 809–814.

[63] Eisenbeis, S., Proffitt, W., Coles, M., Truffault, V., Shanmugaratnam, S., Meiler, J., and Höcker, B. (2012) Potential of fragment recombination for rational design of proteins. *J Am Chem Soc*, **134**, 4019–4022.

[64] Javadi, Y. and Itzhaki, L. S. (2013) Tandem-repeat proteins: regularity plus modularity equals design-ability. *Curr Opin Struct Biol*, **23**, 622–631.

[65] Boersma, Y. L. and Plückthun, A. (2011) Darpins and other repeat protein scaffolds: advances in engineering and applications. *Curr Opin Biotechnol*, **22**, 849–857.

[66] Ferreiro, D. U., Walczak, A. M., Komives, E. A., and Wolynes, P. G. (2008) The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Comput Biol*, **4**, e1000070.

[67] Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K., and Plückthun, A. (2008) Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J Mol Biol*, **376**, 241–257.

[68] Hagai, T., Azia, A., Trizac, E., and Levy, Y. (2012) Modulation of folding kinetics of repeat proteins: interplay between intra- and interdomain interactions. *Biophys J*, **103**, 1555–1565.

[69] Marold, J. D., Kavran, J. M., Bowman, G. D., and Barrick, D. (2015) A naturally occurring repeat protein with high internal sequence identity defines a new class of tpr-like proteins. *Structure*, **23**, 2055–2065.

[70] Broom, A., Ma, S. M., Xia, K., Rafalia, H., Trainor, K., Colón, W., Gosavi, S., and Meiering, E. M. (2015) Designed protein reveals structural determinants of extreme kinetic stability. *Proc Natl Acad Sci U S A*, **112**, 14605–14610.

[71] Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1995) Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J Mol Biol*, **252**, 460–471.

[72] Sawyer, N., Chen, J., and Regan, L. (2013) All repeats are not equal: a module-based approach to guide repeat protein design. *J Mol Biol*, **425**, 1826–1838.

[73] Voet, A. R. D., Noguchi, H., Addy, C., Zhang, K. Y. J., and Tame, J. R. H. (2015) Biomineralization of a cadmium chloride nanocrystal by a designed symmetrical protein. *Angew Chem Int Ed Engl*, **54**, 9857–9860.

[74] Fortenberry, C., Bowman, E. A., Proffitt, W., Dorr, B., Combs, S., Harp, J., Mizoue, L., and Meiler, J. (2011) Exploring symmetry as an avenue to the computational design of large protein domains. *J Am Chem Soc*, **133**, 18026–18029.

[75] Carstensen, L., Sperl, J. M., Bocola, M., List, F., Schmid, F. X., and Sterner, R. (2012) Conservation of the folding mechanism between designed primordial (betaalpha)8-barrel proteins and their modern descendant. *J Am Chem Soc*, **134**, 12786–12791.

[76] Figueroa, M., Oliveira, N., Lejeune, A., Kaufmann, K. W., Dorr, B. M., Matagne, A., Martial, J. A., Meiler, J., and de Weerdt, C. V. (2013) Octarellin vi: using rosetta to design a putative artificial (betaalpha)8 protein. *PLoS One*, **8**, e71858.

[77] Nagarajan, D., Deka, G., and Rao, M. (2015) Design of symmetric tim barrel proteins from first principles. *BMC Biochem*, **16**, 18.

[78] Broom, A., Gosavi, S., and Meiering, E. M. (2015) Protein unfolding rates correlate as strongly as folding rates with native structure. *Protein Sci*, **24**, 580–587.

[79] Lee, J., Blaber, S. I., Dubey, V. K., and Blaber, M. (2011) A polypeptide "building block" for the beta-trefoil fold identified by "top-down symmetric deconstruction". *J Mol Biol*, **407**, 744–763.

[80] Longo, L. M., Lee, J., Tenorio, C. A., and Blaber, M. (2013) Alternative folding nuclei definitions facilitate the evolution of a symmetric protein fold from a smaller peptide motif. *Structure*, **21**, 2042–2050.

[81] Grigoryan, G. and Degrado, W. F. (2011) Probing designability via a generalized model of helical bundle geometry. *J Mol Biol*, **405**, 1079–1100.

[82] Harbury, P. B., Zhang, T., Kim, P. S., and Alber, T. (1993) A switch between two-, three-, and four-stranded coiled coils in gcn4 leucine zipper mutants. *Science*, **262**, 1401–1407.

[83] Gradišar, H., Božič, S., Doles, T., Vengust, D., Hafner-Bratkovič, I., Mertelj, A., Webb, B., Šali, A., Klavžar, S., and Jerala, R. (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat Chem Biol*, **9**, 362–366.

[84] Fletcher, J. M., et al. (2013) Self-assembling cages from coiled-coil peptide modules. *Science*, **340**, 595–599.

[85] Potapov, V., Kaplan, J. B., and Keating, A. E. (2015) Data-driven prediction and design of bzip coiled-coil interactions. *PLoS Comput Biol*, **11**, e1004046.

[86] Pace, C. N., Scholtz, J. M., and Grimsley, G. R. (2014) Forces stabilizing proteins. *FEBS Lett*, **588**, 2177–2184.

[87] Sheffler, W. and Baker, D. (2009) Rosettaholes: rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Sci*, **18**, 229–239.

[88] Dai, L., Yang, Y., Kim, H. R., and Zhou, Y. (2010) Improving computational protein design by using structure-derived sequence profile. *Proteins*, **78**, 2338–2348.

[89] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005) The foldx web server: an online force field. *Nucleic Acids Res*, **33**, W382–W388.

[90] Mitra, P., Shultis, D., Brender, J. R., Czajka, J., Marsh, D., Gray, F., Cierpicki, T., and Zhang, Y. (2013) An evolution-based approach to de novo protein design and case study on mycobacterium tuberculosis. *PLoS Comput Biol*, **9**, e1003298.

[91] Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., and Liu, H. (2014) Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat Commun*, **5**, 5330.

[92] Potapov, V., Cohen, M., and Schreiber, G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel*, **22**, 553–560.

[93] Khan, S. and Vihinen, M. (2010) Performance of protein stability predictors. *Hum Mutat*, **31**, 675–684.

[94] Das, R. (2011) Four small puzzles that rosetta doesn't solve. *PLoS One*, **6**, e20044.

[95] Murphy, G. S., Sathyamoorthy, B., Der, B. S., Machius, M. C., Pulavarti, S. V., Szyperski, T., and Kuhlman, B. (2015) Computational de novo design of a four-helix bundle protein–dnd-4hb. *Protein Sci*, **24**, 434–445.

[96] Ollikainen, N. and Kortemme, T. (2013) Computational protein design quantifies structural constraints on amino acid covariation. *PLoS Comput Biol*, **9**, e1003313.

[97] Zhang, Z. and Chan, H. S. (2010) Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proc Natl Acad Sci U S A*, **107**, 2920–2925.

[98] Yadahalli, S. and Gosavi, S. (2014) Designing cooperativity into the designed protein top7. *Proteins*, **82**, 364–374.

[99] Truong, H. H., Kim, B. L., Schafer, N. P., and Wolynes, P. G. (2013) Funneling and frustration in the energy landscapes of some designed and simplified proteins. *J Chem Phys*, **139**, 121908.

[100] Tzul, F. O., Schweiker, K. L., and Makhatadze, G. I. (2015) Modulation of folding energy landscape by charge-charge interactions: linking experiments with computational modeling. *Proc Natl Acad Sci U S A*, **112**, E259–E266.

[101] Chung, H. S., Piana-Agostinetti, S., Shaw, D. E., and Eaton, W. A. (2015) Structural origin of slow diffusion in protein folding. *Science*, **349**, 1504–1510.

[102] Rao, V. H. G. and Gosavi, S. (2016) Using the folding landscapes of proteins to understand protein function. *Curr Opin Struct Biol*, **36**, 67–74.

[103] Mou, Y., Huang, P.-S., Thomas, L. M., and Mayo, S. L. (2015) Using molecular dynamics simulations as an aid in the prediction of domain swapping of computationally designed protein variants. *J Mol Biol*, **427**, 2697–2706.

[104] Jacak, R., Leaver-Fay, A., and Kuhlman, B. (2012) Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins*, **80**, 825–838.

[105] Der, B. S., Kluwe, C., Miklos, A. E., Jacak, R., Lyskov, S., Gray, J. J., Georgiou, G., Ellington, A. D., and Kuhlman, B. (2013) Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. *PLoS One*, **8**, e64363.

[106] Suezaki, Y. and Go, N. (1975) Breathing mode of conformational fluctuations in globular proteins. *Int. J. Peptide Protein Res.*, **7**, 333–334.

[107] Shapovalov, M. V. and Dunbrack, R. L. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**, 844–858.

[108] Durrant, J. D. and McCammon, J. A. (2011) Molecular dynamics simulations and drug discovery. *BMC Biol*, **9**, 71.

[109] Bucher, D. and Rothlisberger, U. (2010) Molecular simulations of ion channels: a quantum chemist's perspective. *J Gen Physiol*, **135**, 549–554.

[110] Piana, S., Klepeis, J. L., and Shaw, D. E. (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*, **24**, 98–105.

[111] Miao, Y., Feixas, F., Eun, C., and McCammon, J. A. (2015) Accelerated molecular dynamics simulations of protein folding. *J Comput Chem*, **36**, 1536–1549.

[112] Rauscher, S. and Pomès, R. (2010) Molecular simulations of protein disorder. *Biochem Cell Biol*, **88**, 269–290.

[113] Redler, R. L., Shirvanyants, D., Dagliyan, O., Ding, F., Kim, D. N., Kota, P., Proctor, E. A., Ramachandran, S., Tandon, A., and Dokholyan, N. V. (2014) Computational approaches to understanding protein aggregation in neurodegeneration. *J Mol Cell Biol*, **6**, 104–115.

[114] Avila, C. L., Drechsel, N. J. D., Alcántara, R., and Villà-Freixa, J. (2011) Multiscale molecular dynamics of protein aggregation. *Curr Protein Pept Sci*, **12**, 221–234.

[115] Senn, H. M. and Thiel, W. (2007) Qm/mm studies of enzymes. *Curr Opin Chem Biol*, **11**, 182–187.

[116] Lupas, A. N., Ponting, C. P., and Russell, R. B. (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*, **134**, 191–203.

[117] Eisenbeis, S. and Höcker, B. (2010) Evolutionary mechanism as a template for protein engineering. *J Pept Sci*, **16**, 538–544.

[118] Houbrechts, A., Moreau, B., Abagyan, R., Mainfroid, V., Préaux, G., Lamproye, A., Poncin, A., Goormaghtigh, E., Ruysschaert, J. M., and Martial, J. A. (1995) Second-generation octarellins: two new de novo (beta/alpha)8 polypeptides designed for investigating the influence of beta-residue packing on the alpha/beta-barrel structure stability. *Protein Eng*, **8**, 249–259.

[119] Höcker, B., Lochner, A., Seitz, T., Claren, J., and Sterner, R. (2009) High-resolution crystal structure of an artificial (betaalpha)(8)-barrel protein designed from identical half-barrels. *Biochemistry*, **48**, 1145–1147.

[120] Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000) Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.

[121] Hill, R. B., Raleigh, D. P., Lombardi, A., and DeGrado, W. F. (2000) De novo design of helical bundles as models for understanding protein folding and function. *Acc Chem Res*, **33**, 745–754.

[122] Ponting, C. P. and Russell, R. B. (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J Mol Biol*, **302**, 1041–1047.

[123] Murzin, A. G., Lesk, A. M., and Chothia, C. (1992) beta-trefoil fold. patterns of structure and sequence in the kunitz inhibitors interleukins-1 beta and 1 alpha and fibroblast growth factors. *J Mol Biol*, **223**, 531–543.

[124] Finn, R. D., et al. (2010) The pfam protein families database. *Nucleic Acids Res*, **38**, D211–D222.

[125] Mukhopadhyay, D. (2000) The molecular evolutionary history of a winged bean alpha-chymotrypsin inhibitor and modeling of its mutations through structural analyses. *J Mol Evol*, **50**, 214–223.

[126] McLachlan, A. D. (1979) Three-fold structural pattern in the soybean trypsin inhibitor (kunitz). *J Mol Biol*, **133**, 557–563.

[127] Lee, J. and Blaber, M. (2011) Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc Natl Acad Sci U S A*, **108**, 126–130.

[128] Houliston, R. S., Liu, C., Singh, L. M. R., and Meiering, E. M. (2002) ph and urea dependence of amide hydrogen-deuterium exchange rates in the beta-trefoil protein hisactophilin. *Biochemistry*, **41**, 1182–1194.

[129] Dubey, V. K., Lee, J., Somasundaram, T., Blaber, S., and Blaber, M. (2007) Spackling the crack: stabilizing human fibroblast growth factor-1 by targeting the n and c terminus beta-strand interactions. *J Mol Biol*, **371**, 256–268.

[130] Capraro, D. T., Roy, M., Onuchic, J. N., and Jennings, P. A. (2008) Backtracking on the folding landscape of the beta-trefoil protein interleukin-1beta? *Proc Natl Acad Sci U S A*, **105**, 14844–14848.

[131] Grahn, E., Askarieh, G., Holmner, A., Tateno, H., Winter, H. C., Goldstein, I. J., and Krengel, U. (2007) Crystal structure of the marasmius oreades mushroom lectin in complex with a xenotransplantation epitope. *J Mol Biol*, **369**, 710–721.

[132] Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, **60**, 2256–2268.

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

[134] Marchler-Bauer, A., et al. (2009) Cdd: specific functional annotation with the conserved domain database. *Nucleic Acids Res*, **37**, D205–D210.

[135] Orengo, C. A. and Thornton, J. M. (2005) Protein families and their evolution-a structural perspective. *Annu Rev Biochem*, **74**, 867–900.

[136] Chaudhuri, I., Söding, J., and Lupas, A. N. (2008) Evolution of the beta-propeller fold. *Proteins*, **71**, 795–803.

[137] Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2008) Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design. *Structure*, **16**, 1799–1805.

[138] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

[139] Hazes, B. (1996) The (qxw)3 domain: a flexible lectin scaffold. *Protein Sci*, **5**, 1490–1501.

[140] Stenmark, P., Gurmu, D., and Nordlund, P. (2004) Crystal structure of caib, a type-iii coa transferase in carnitine metabolism. *Biochemistry*, **43**, 13996–14003.

[141] Winter, H. C., Mostafapour, K., and Goldstein, I. J. (2002) The mushroom marasmius oreades lectin is a blood group type b agglutinin that recognizes the galalpha 1,3gal and galalpha 1,3galbeta 1,4glcnac porcine xenotransplantation epitopes with high affinity. *J Biol Chem*, **277**, 14996–15001.

[142] Goodsell, D. S. and Olson, A. J. (2000) Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, **29**, 105–153.

[143] Vivian, J. T. and Callis, P. R. (2001) Mechanisms of tryptophan fluorescence shifts in proteins. *Biophys J*, **80**, 2093–2109.

[144] Akanuma, S., Matsuba, T., Ueno, E., Umeda, N., and Yamagishi, A. (2010) Mimicking the evolution of a thermally stable monomeric four-helix bundle by fusion of four identical single-helix peptides. *J Biochem*, **147**, 371–379.

[145] Zimm, B. (1948) The scattering of light and the radial distribution function of high polymer solutions. *J. Chem. Phys.*, **16**, 1093–1099.

[146] Osterhout, J. J., Handel, T., Na, G., Toumadje, A., Long, R. C., Connolly, P. J., Hoch, J. C., Johnson, C., Live, D., and DeGrado, W. F. (1992) Characterization of the structural properties of alpha-1b, a peptide designed to form a four-helix bundle. *J. Am. Chem. Soc.*, **114**, 331–337.

[147] Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85–94.

[148] Main, E. R. G., Lowe, A. R., Mochrie, S. G. J., Jackson, S. E., and Regan, L. (2005) A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol*, **15**, 464–471.

[149] Yadid, I., Kirshenbaum, N., Sharon, M., Dym, O., and Tawfik, D. S. (2010) Metamorphic proteins mediate evolutionary transitions of structure. *Proc Natl Acad Sci U S A*, **107**, 7287–7292.

[150] Barrick, D., Ferreiro, D. U., and Komives, E. A. (2008) Folding landscapes of ankyrin repeat proteins: experiments meet theory. *Curr Opin Struct Biol*, **18**, 27–34.

[151] André, I., Strauss, C. E. M., Kaplan, D. B., Bradley, P., and Baker, D. (2008) Emergence of symmetry in homooligomeric biological assemblies. *Proc Natl Acad Sci U S A*, **105**, 16148–16152.

[152] Beisel, H. G., Kawabata, S., Iwanaga, S., Huber, R., and Bode, W. (1999) Tachylectin-2: crystal structure of a specific glcnac/galnac-binding lectin involved in the innate immunity host defense of the japanese horseshoe crab tachypleus tridentatus. *EMBO J*, **18**, 2313–2322.

[153] Tokuriki, N. and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*, **19**, 596–604.

[154] Meiering, E. M., Serrano, L., and Fersht, A. R. (1992) Effect of active site residues in barnase on activity and stability. *J Mol Biol*, **225**, 585–589.

[155] Li, H., Helling, R., Tang, C., and Wingreen, N. (1996) Emergence of preferred structures in a simple model of protein folding. *Science*, **273**, 666–669.

[156] Miyanaga, A., Koseki, T., Matsuzawa, H., Wakagi, T., Shoun, H., and Fushinobu, S. (2004) Crystal structure of a family 54 alpha-l-arabinofuranosidase reveals a novel carbohydrate-binding module that can bind arabinose. *J Biol Chem*, **279**, 44907–44914.

[157] Sedeh, R. S., Fedorov, A. A., Fedorov, E. V., Ono, S., Matsumura, F., Almo, S. C., and Bathe, M. (2010) Structure, evolutionary conservation, and conformational dynamics of homo sapiens fascin-1, an f-actin crosslinking protein. *J Mol Biol*, **400**, 589–604.

[158] Collins, B. E. and Paulson, J. C. (2004) Cell surface biology mediated by low affinity multivalent protein-glycan interactions. *Curr Opin Chem Biol*, **8**, 617–625.

[159] Doxey, A. C., Lynch, M. D. J., Müller, K. M., Meiering, E. M., and McConkey, B. J. (2008) Insights into the evolutionary origins of clostridial neurotoxins from analysis of the clostridium botulinum strain a neurotoxin gene cluster. *BMC Evol Biol*, **8**, 316.

[160] Fernandez-Fuentes, N., Dybas, J. M., and Fiser, A. (2010) Structural characteristics of novel protein folds. *PLoS Comput Biol*, **6**, e1000750.

[161] Yadid, I. and Tawfik, D. S. (2011) Functional beta-propeller lectins by tandem duplications of repetitive units. *Protein Eng Des Sel*, **24**, 185–195.

[162] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–540.

[163] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Res*, **28**, 235–242.

[164] Gouy, M., Guindon, S., and Gascuel, O. (2010) Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*, **27**, 221–224.

[165] Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779–815.

[166] Notenboom, V., Boraston, A. B., Williams, S. J., Kilburn, D. G., and Rose, D. R. (2002) High-resolution crystal structures of the lectin-like xylan binding domain from streptomyces lividans xylanase 10a with bound substrates reveal a novel mode of xylan binding. *Biochemistry*, **41**, 4246–4254.

[167] Arndt, J. W., Gu, J., Jaroszewski, L., Schwarzenbacher, R., Hanson, M. A., Lebeda, F. J., and Stevens, R. C. (2005) The structure of the neurotoxin-associated protein

ha33/a from clostridium botulinum suggests a reoccurring beta-trefoil fold in the progenitor toxin complex. *J Mol Biol*, **346**, 1083–1093.

[168] Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) How to measure and predict the molar absorption coefficient of a protein. *Protein Sci*, **4**, 2411–2423.

[169] Pflugrath, J. W. (1999) The finer things in x-ray diffraction data collection. *Acta Crystallogr D Biol Crystallogr*, **55**, 1718–1725.

[170] Long, F., Vagin, A. A., Young, P., and Murshudov, G. N. (2008) Balbes: a molecular-replacement pipeline. *Acta Crystallogr D Biol Crystallogr*, **64**, 125–132.

[171] Adams, P. D., et al. (2010) Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*, **66**, 213–221.

[172] Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of coot. *Acta Crystallogr D Biol Crystallogr*, **66**, 486–501.

[173] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, **66**, 12–21.

[174] Laskowski, R., MacArthur, M., Moss, D., and Thornton, J. (1993) Procheck - a program to check the stereochemical quality of protein structures. *J. App. Cyrst.*, **26**, 283–291.

[175] Painter, J. and Merritt, E. A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing tls motion. *Acta Crystallogr D Biol Crystallogr*, **62**, 439–450.

[176] Hwang, T. and Olson, A. (1995) Water suppression that works: excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J. Magn. Reson. A*, **112**, 275–279.

[177] Blixt, O., et al. (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci U S A*, **101**, 17033–17038.

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

[179] Alva, V., Söding, J., and Lupas, A. N. (2015) A vocabulary of ancient peptides at the origin of folded proteins. *Elife*, **4**, e09410.

[180] Kopec, K. O. and Lupas, A. N. (2013) Beta-propeller blades as ancestral peptides in protein evolution. *PLoS One*, **8**, e77074.

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

[182] Shimizu, T., Vassylyev, D. G., Kido, S., Doi, Y., and Morikawa, K. (1994) Crystal structure of vitelline membrane outer layer protein i (vmo-i): a folding motif with homologous greek key structures related by an internal three-fold symmetry. *EMBO J*, **13**, 1003–1010.

[183] Sharma, A., Chandran, D., Singh, D. D., and Vijayan, M. (2007) Multiplicity of carbohydrate-binding sites in beta-prism fold lectins: occurrence and possible evolutionary implications. *J Biosci*, **32**, 1089–1110.

[184] Pigott, C. R. and Ellar, D. J. (2007) Role of receptors in bacillus thuringiensis crystal toxin activity. *Microbiol Mol Biol Rev*, **71**, 255–281.

[185] Ziółkowska, N. E., O'Keefe, B. R., Mori, T., Zhu, C., Giomarelli, B., Vojdani, F., Palmer, K. E., McMahon, J. B., and Wlodawer, A. (2006) Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding. *Structure*, **14**, 1127–1135.

[186] Swanson, M. D., et al. (2015) Engineering a therapeutic lectin by uncoupling mitogenicity from antiviral activity. *Cell*, **163**, 746–758.

[187] Sankaranarayanan, R., Sekar, K., Banerjee, R., Sharma, V., Surolia, A., and Vijayan, M. (1996) A novel mode of carbohydrate recognition in jacalin, a moraceae plant lectin with a beta-prism fold. *Nat Struct Biol*, **3**, 596–603.

[188] Hoorelbeke, B., Huskens, D., Férir, G., François, K. O., Takahashi, A., Laethem, K. V., Schols, D., Tanaka, H., and Balzarini, J. (2010) Actinohivin, a broadly neutralizing prokaryotic lectin, inhibits hiv-1 infection by specifically targeting high-mannose-type glycans on the gp120 envelope. *Antimicrob Agents Chemother*, **54**, 3287–3301.

[189] Wang, Z., Chen, Z., Yang, Q., Jiang, Y., Lin, L., Liu, X., and Wu, K. (2014) Vitelline membrane outer layer 1 homolog interacts with lysozyme c and promotes the stabilization of tear film. *Invest Ophthalmol Vis Sci*, **55**, 6722–6727.

[190] Mitchell, M. (1998) *An Introduction to Genetic Algorithms.* MIT Press.

[191] Xu, H. and Shaw, D. E. (2016) A simple model of multivalent adhesion and its application to influenza infection. *Biophys J*, **110**, 218–233.

[192] Daggett, V. and Fersht, A. (2003) The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol*, **4**, 497–502.

[193] Sosnick, T. R. and Barrick, D. (2011) The folding of single domain proteins–have we reached a consensus? *Curr Opin Struct Biol*, **21**, 12–24.

[194] Chan, H. S., Zhang, Z., Wallin, S., and Liu, Z. (2011) Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu Rev Phys Chem*, **62**, 301–326.

[195] Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. (2008) The protein folding problem. *Annu Rev Biophys*, **37**, 289–316.

[196] Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*, **48**, 545–600.

[197] Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, **277**, 985–994.

[198] Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Sci*, **12**, 2057–2062.

[199] Lane, T. J. and Pande, V. S. (2013) Inferring the rate-length law of protein folding. *PLoS One*, **8**, e78606.

[200] Gutin, Abkevich, and Shakhnovich (1996) Chain length scaling of protein folding time. *Phys Rev Lett*, **77**, 5433–5436.

[201] Zwanzig, R., Szabo, A., and Bagchi, B. (1992) Levinthal's paradox. *Proc Natl Acad Sci U S A*, **89**, 20–22.

[202] Li, M., Klimov, D., and Thirumalali, D. (2002) Dependence of folding rates on protein length. *J. Phys. Chem. B.*, **106**, 8302–8305.

[203] Wolynes, P. G. (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci U S A*, **94**, 6170–6175.

[204] Naganathan, A. N. and Muñoz, V. (2005) Scaling of folding times with protein size. *J Am Chem Soc*, **127**, 480–481.

[205] Istomin, A. Y., Jacobs, D. J., and Livesay, D. R. (2007) On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. *Protein Sci*, **16**, 2564–2569.

[206] Rollins, G. C. and Dill, K. A. (2014) General mechanism of two-state protein folding kinetics. *J Am Chem Soc*, **136**, 11420–11427.

[207] Paci, E., Lindorff-Larsen, K., Dobson, C. M., Karplus, M., and Vendruscolo, M. (2005) Transition state contact orders correlate with protein folding rates. *J Mol Biol*, **352**, 495–500.

[208] Faísca, P. F. N., Travasso, R. D. M., Parisi, A., and Rey, A. (2012) Why do protein folding rates correlate with metrics of native topology? *PLoS One*, **7**, e35599.

[209] Sosnick, T. R. (2008) Kinetic barriers and the role of topology in protein and rna folding. *Protein Sci*, **17**, 1308–1318.

[210] Gromiha, M. M. and Selvaraj, S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol*, **310**, 27–32.

[211] Zou, T. and Ozkan, S. B. (2011) Local and non-local native topologies reveal the underlying folding landscape of proteins. *Phys Biol*, **8**, 066011.

[212] Zhou, H. and Zhou, Y. (2002) Folding rate prediction using total contact distance. *Biophys J*, **82**, 458–463.

[213] Ouyang, Z. and Liang, J. (2008) Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci*, **17**, 1256–1263.

[214] Micheletti, C. (2003) Prediction of folding rates and transition-state placement from native-state geometry. *Proteins*, **51**, 74–84.

[215] Tejera, E., Machado, A., Rebelo, I., and Nieto-Villar, J. (2009) Fractal protein structure revisited: topological, kinetic and thermodynamic relationships. *Phys. A Stat. Mech. Appl.*, **388**, 4600–4608.

[216] Rustad, M. and Ghosh, K. (2012) Why and how does native topology dictate the folding speed of a protein? *J Chem Phys*, **137**, 205104.

[217] Su, J. G., Li, C. H., Hao, R., Chen, W. Z., and Wang, C. X. (2008) Protein unfolding behavior studied by elastic network model. *Biophys J*, **94**, 4586–4596.

[218] Jung, J., Buglass, A., and Lee, E.-K. (2010) Topological quantities determining the folding/unfolding rate of two-state folding proteins. *J. Solut. Chem.*, **39**, 943–958.

[219] Jung, J., Lee, J., and Moon, H.-T. (2005) Topological determinants of protein unfolding rates. *Proteins*, **58**, 389–395.

[220] Harihar, B. and Selvaraj, S. (2011) Application of long-range order to predict unfolding rates of two-state proteins. *Proteins*, **79**, 880–887.

[221] Maxwell, K. L., et al. (2005) Protein folding: defining a “standard” set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci*, **14**, 602–616.

[222] Wensley, B. G., Gärtner, M., Choo, W. X., Batey, S., and Clarke, J. (2009) Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J Mol Biol*, **390**, 1074–1085.

[223] Ferguson, N., Sharpe, T. D., Schartau, P. J., Sato, S., Allen, M. D., Johnson, C. M., Rutherford, T. J., and Fersht, A. R. (2005) Ultra-fast barrier-limited folding in the peripheral subunit-binding domain family. *J Mol Biol*, **353**, 427–446.

[224] Perl, D., Welker, C., Schindler, T., Schröder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat Struct Biol*, **5**, 229–235.

[225] Garbuzynskiy, S. O., Ivankov, D. N., Bogatyreva, N. S., and Finkelstein, A. V. (2013) Golden triangle for folding rates of globular proteins. *Proc Natl Acad Sci U S A*, **110**, 147–150.

[226] Chavez, L. L., Onuchic, J. N., and Clementi, C. (2004) Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc*, **126**, 8426–8432.

[227] Plaxco, K. W., Simons, K. T., Ruczinski, I., and Baker, D. (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry*, **39**, 11177–11183.

[228] Gráczer, E., Varga, A., Hajdú, I., Melnik, B., Szilágyi, A., Semisotnov, G., Závodszky, P., and Vas, M. (2007) Rates of unfolding, rather than refolding, determine thermal stabilities of thermophilic, mesophilic, and psychrotrophic 3-isopropylmalate dehydrogenases. *Biochemistry*, **46**, 11536–11549.

[229] Dobson, C. M. (2003) Protein folding and misfolding. *Nature*, **426**, 884–890.

[230] Matouschek, A. (2003) Protein unfolding–an important process in vivo? *Curr Opin Struct Biol*, **13**, 98–109.

[231] Sancho, D. D. and Muñoz, V. (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys*, **13**, 17030–17043.

[232] Fiebig, K. and Dill, K. (1993) Protein core assembly processes. *J. Chem. Phys.*, **98**, 3475.

[233] Nickson, A. A. and Clarke, J. (2010) What lessons can be learned from studying the folding of homologous proteins? *Methods*, **52**, 38–50.

[234] Jackson, S. E. (1998) How do small single-domain proteins fold? *Fold Des*, **3**, R81–R91.

[235] Naganathan, A. N. and Muñoz, V. (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc Natl Acad Sci U S A*, **107**, 8611–8616.

[236] Gosavi, S. (2013) Understanding the folding-function tradeoff in proteins. *PLoS One*, **8**, e61222.

[237] Yi, Q., Scalley-Kim, M. L., Alm, E. J., and Baker, D. (2000) Nmr characterization of residual structure in the denatured state of protein l. *J Mol Biol*, **299**, 1341–1351.

[238] Ratcliff, K. and Marqusee, S. (2010) Identification of residual structure in the unfolded state of ribonuclease h1 from the moderately thermophilic chlorobium tepidum: comparison with thermophilic and mesophilic homologues. *Biochemistry*, **49**, 5167–5175.

[239] Zarrine-Afsar, A., Wallin, S., Neculai, A. M., Neudecker, P., Howell, P. L., Davidson, A. R., and Chan, H. S. (2008) Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding. *Proc Natl Acad Sci U S A*, **105**, 9999–10004.

[240] Shental-Bechor, D., Smith, M. T. J., Mackenzie, D., Broom, A., Marcovitz, A., Ghashut, F., Go, C., Bralha, F., Meiering, E. M., and Levy, Y. (2012) Nonnative interactions regulate folding and switching of myristoylated protein. *Proc Natl Acad Sci U S A*, **109**, 17839–17844.

[241] Clementi, C. and Plotkin, S. S. (2004) The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci*, **13**, 1750–1766.

[242] Machius, M., Declerck, N., Huber, R., and Wiegand, G. (2003) Kinetic stabilization of bacillus licheniformis alpha-amylase through introduction of hydrophobic residues at the surface. *J Biol Chem*, **278**, 11546–11553.

[243] Cavagnero, S., Debe, D. A., Zhou, Z. H., Adams, M. W., and Chan, S. I. (1998) Kinetic role of electrostatic interactions in the unfolding of hyperthermophilic and mesophilic rubredoxins. *Biochemistry*, **37**, 3369–3376.

[244] Rodriguez-Larrea, D., Minning, S., Borchert, T. V., and Sanchez-Ruiz, J. M. (2006) Role of solvation barriers in protein kinetic stability. *J Mol Biol*, **360**, 715–724.

[245] Manning, M. and Colón, W. (2004) Structural basis of protein kinetic stability: resistance to sodium dodecyl sulfate suggests a central role for rigidity and a bias toward beta-sheet structure. *Biochemistry*, **43**, 11248–11254.

[246] Sanchez-Ruiz, J. M. (2010) Protein kinetic stability. *Biophys Chem*, **148**, 1–15.

[247] Lawrence, C., Kuge, J., Ahmad, K., and Plaxco, K. W. (2010) Investigation of an anomalously accelerating substitution in the folding of a prototypical two-state protein. *J Mol Biol*, **403**, 446–458.

[248] Bogatyreva, N. S., Osypov, A. A., and Ivankov, D. N. (2009) Kineticdb: a database of protein folding kinetics. *Nucleic Acids Res*, **37**, D342–D346.

[249] Smith, M. T. J., Meissner, J., Esmonde, S., Wong, H. J., and Meiering, E. M. (2010) Energetics and mechanisms of folding and flipping the myristoyl switch in the beta-trefoil protein, hisactophilin. *Proc Natl Acad Sci U S A*, **107**, 20952–20957.

[250] Spector, S. and Raleigh, D. P. (1999) Submillisecond folding of the peripheral subunit-binding domain. *J Mol Biol*, **293**, 763–768.

[251] Koga, N. and Takada, S. (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. *J Mol Biol*, **313**, 171–180.

[252] Kouza, M., Li, M. S., O'brien, E. P., Hu, C.-K., and Thirumalai, D. (2006) Effect of finite size on cooperativity and rates of protein folding. *J Phys Chem A*, **110**, 671–676.

[253] Longo, L. M., Kumru, O. S., Middaugh, C. R., and Blaber, M. (2014) Evolution and design of protein structure by folding nucleus symmetric expansion. *Structure*, **22**, 1377–1384.

[254] Rajagopalan, S., et al. (2014) Design of activated serine-containing catalytic triads with atomic-level accuracy. *Nat Chem Biol*, **10**, 386–391.

[255] Korendovych, I. V. and DeGrado, W. F. (2014) Catalytic efficiency of designed catalytic proteins. *Curr Opin Struct Biol*, **27**, 113–121.

[256] Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995) A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A*, **92**, 452–456.

[257] Capraro, D. T., Roy, M., Onuchic, J. N., Gosavi, S., and Jennings, P. A. (2012) beta-bulge triggers route-switching on the functional landscape of interleukin-1beta. *Proc Natl Acad Sci U S A*, **109**, 1490–1493.

[258] Ferreiro, D. U., Komives, E. A., and Wolynes, P. G. (2014) Frustration in biomolecules. *Q Rev Biophys*, **47**, 285–363.

[259] Gershenson, A., Gierasch, L. M., Pastore, A., and Radford, S. E. (2014) Energy landscapes of functional proteins are inherently risky. *Nat Chem Biol*, **10**, 884–891.

[260] Leaver-Fay, A., et al. (2011) Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, **487**, 545–574.

[261] Nymeyer, H., García, A. E., and Onuchic, J. N. (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc Natl Acad Sci U S A*, **95**, 5921–5928.

[262] Hyeon, C. and Thirumalai, D. (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat Commun*, **2**, 487.

[263] Borgia, M. B., Borgia, A., Best, R. B., Steward, A., Nettels, D., Wunderlich, B., Schuler, B., and Clarke, J. (2011) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature*, **474**, 662–665.

[264] Jaswal, S. S., Sohl, J. L., Davis, J. H., and Agard, D. A. (2002) Energetic landscape of alpha-lytic protease optimizes longevity through kinetic stability. *Nature*, **415**, 343–346.

[265] Javadi, Y. and Main, E. R. G. (2009) Exploring the folding energy landscape of a series of designed consensus tetratricopeptide repeat proteins. *Proc Natl Acad Sci U S A*, **106**, 17383–17388.

[266] Sancho, J., Meiering, E. M., and Fersht, A. R. (1991) Mapping transition states of protein unfolding by protein engineering of ligand-binding sites. *J Mol Biol*, **221**, 1007–1014.

[267] Wittung-Stafshede, P. (2002) Role of cofactors in protein folding. *Acc Chem Res*, **35**, 201–208.

[268] Stigler, J. and Rief, M. (2012) Calcium-dependent folding of single calmodulin molecules. *Proc Natl Acad Sci U S A*, **109**, 17814–17819.

[269] Liu, P.-F. and Park, C. (2012) Selective stabilization of a partially unfolded protein by a metabolite. *J Mol Biol*, **422**, 403–413.

[270] Thirumalai, D. (1995) From minimal models to real proteins - time scales for protein folding kinetics. *J. Phys. I*, **5**, 1457–1467.

[271] Park, C., Zhou, S., Gilmore, J., and Marqusee, S. (2007) Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. *J Mol Biol*, **368**, 1426–1437.

[272] Xia, K., Manning, M., Hesham, H., Lin, Q., Bystroff, C., and Colón, W. (2007) Identifying the subproteome of kinetically stable proteins via diagonal 2d sds/page. *Proc Natl Acad Sci U S A*, **104**, 17329–17334.

[273] Xia, K., Zhang, S., Bathrick, B., Liu, S., Garcia, Y., and Colón, W. (2012) Quantifying the kinetic stability of hyperstable proteins via time-dependent sds trapping. *Biochemistry*, **51**, 100–107.

[274] Best, R. B., Hummer, G., and Eaton, W. A. (2013) Native contacts determine protein folding mechanisms in atomistic simulations. *Proc Natl Acad Sci U S A*, **110**, 17874–17879.

[275] Cota, E. and Clarke, J. (2000) Folding of beta-sandwich proteins: three-state transition of a fibronectin type iii module. *Protein Sci*, **9**, 112–120.

[276] Geierhaas, C. D., Nickson, A. A., Lindorff-Larsen, K., Clarke, J., and Vendruscolo, M. (2007) Bppred: a web-based computational tool for predicting biophysical parameters of proteins. *Protein Sci*, **16**, 125–134.

[277] Whitmore, L. and Wallace, B. A. (2004) Dichroweb, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res*, **32**, W668–W673.

[278] Whitmore, L. and Wallace, B. A. (2008) Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, **89**, 392–400.

[279] Johnson, W. C. (1999) Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins*, **35**, 307–312.

[280] Semisotnov, G. V., Rodionova, N. A., Razgulyaev, O. I., Uversky, V. N., Gripas', A. F., and Gilmanshin, R. I. (1991) Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers*, **31**, 119–128.

[281] Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with tm-score = 0.5? *Bioinformatics*, **26**, 889–895.

[282] Clementi, C., Nymeyer, H., and Onuchic, J. N. (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *J Mol Biol*, **298**, 937–953.

[283] Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., and Edelman, M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.

[284] Roberts, E., Eargle, J., Wright, D., and Luthey-Schulten, Z. (2006) Multiseq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, **7**, 382.

[285] Humphrey, W., Dalke, A., and Schulten, K. (1996) Vmd: visual molecular dynamics. *J Mol Graph*, **14**, 33–8, 27–8.

[286] Yang, S., Cho, S. S., Levy, Y., Cheung, M. S., Levine, H., Wolynes, P. G., and Onuchic, J. N. (2004) Domain swapping is a consequence of minimal frustration. *Proc Natl Acad Sci U S A*, **101**, 13786–13791.

[287] Hagai, T. and Levy, Y. (2008) Folding of elongated proteins: conventional or anomalous? *J Am Chem Soc*, **130**, 14253–14262.

[288] Whitford, P. C., Miyashita, O., Levy, Y., and Onuchic, J. N. (2007) Conformational transitions of adenylate kinase: switching by cracking. *J Mol Biol*, **366**, 1661–1671.

[289] Gosavi, S., Chavez, L. L., Jennings, P. A., and Onuchic, J. N. (2006) Topological frustration and the folding of interleukin-1 beta. *J Mol Biol*, **357**, 986–996.

[290] Schimming, S., Schwarz, W. H., and Staudenbauer, W. L. (1991) Properties of a thermoactive beta-1,3-1,4-glucanase (lichenase) from clostridium thermocellum expressed in escherichia coli. *Biochem Biophys Res Commun*, **177**, 447–452.

[291] Saeed, I. A. and Ashraf, S. S. (2009) Denaturation studies reveal significant differences between gfp and blue fluorescent protein. *Int J Biol Macromol*, **45**, 236–241.

[292] Tarentino, A. L., Trimble, R. B., and Maley, F. (1978) endo-beta-n-acetylglucosaminidase from streptomyces plicatus. *Methods Enzymol*, **50**, 574–580.

[293] Zhang, S., Xia, K., Chung, W. K., Cramer, S. M., and Colón, W. (2010) Identifying kinetically stable proteins with capillary electrophoresis. *Protein Sci*, **19**, 888–892.

[294] Liu, C.-L., et al. (2009) Isolation and identification of two novel sds-resistant secreted chitinases from aeromonas schubertii. *Biotechnol Prog*, **25**, 124–131.

[295] Kelch, B. A., Eagen, K. P., Erciyas, F. P., Humphris, E. L., Thomason, A. R., Mitsuiki, S., and Agard, D. A. (2007) Structural and mechanistic exploration of acid resistance: kinetic stability facilitates evolution of extremophilic behavior. *J Mol Biol*, **368**, 870–883.

[296] Hilz, H., Wiegers, U., and Adamietz, P. (1975) Stimulation of proteinase k action by denaturing agents: application to the isolation of nucleic acids and the degradation of 'masked' proteins. *Eur J Biochem*, **56**, 103–108.

[297] Nelson, C. A. (1971) The binding of detergents to proteins. i. the maximum amount of dodecyl sulfate bound to proteins and the resistance to binding of several proteins. *J Biol Chem*, **246**, 3895–3901.

[298] Xu, Y., Xu, D., Gabow, H. N., and Gabow, H. (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.

[299] Bommarius, A. S. and Paye, M. F. (2013) Stabilizing biocatalysts. *Chem Soc Rev*, **42**, 6534–6565.

[300] Silverman, J., et al. (2005) Multivalent avimer proteins evolved by exon shuffling of a family of human receptor domains. *Nat Biotechnol*, **23**, 1556–1561.

[301] Sauerborn, M., Brinks, V., Jiskoot, W., and Schellekens, H. (2010) Immunological mechanism underlying the immune response to recombinant human protein therapeutics. *Trends Pharmacol Sci*, **31**, 53–59.

[302] Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W., and Katayama, D. S. (2010) Stability of protein pharmaceuticals: an update. *Pharm Res*, **27**, 544–575.

[303] Boulet-Audet, M., Byrne, B., and Kazarian, S. G. (2014) High-throughput thermal stability analysis of a monoclonal antibody by attenuated total reflection ft-ir spectroscopic imaging. *Anal Chem*, **86**, 9786–9793.

[304] Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A*, **103**, 5869–5874.

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

[306] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004) Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res*, **32**, D120–D121.

[307] Topham, C. M., Srinivasan, N., and Blundell, T. L. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng*, **10**, 7–21.

[308] Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, **320**, 369–387.

[309] Capriotti, E., Fariselli, P., and Casadio, R. (2005) I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, **33**, W306–W310.

[310] Pokala, N. and Handel, T. M. (2004) Energy functions for protein design i: efficient and accurate continuum electrostatics and solvation. *Protein Sci*, **13**, 925–936.

[311] Parthiban, V., Gromiha, M. M., and Schomburg, D. (2006) Cupsat: prediction of protein stability upon point mutations. *Nucleic Acids Res*, **34**, W239–W242.

[312] Cheng, J., Randall, A., and Baldi, P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.

[313] Yin, S., Ding, F., and Dokholyan, N. V. (2007) Eris: an automated estimator of protein stability. *Nat Methods*, **4**, 466–467.

[314] Deutsch, C. and Krishnamoorthy, B. (2007) Four-body scoring function for mutagenesis. *Bioinformatics*, **23**, 3009–3015.

[315] Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008) A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, **9 Suppl 2**, S6.

[316] Yang, Y. and Zhou, Y. (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci*, **17**, 1212–1219.

[317] Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A., and Böckmann, R. A. (2009) Predicting free energy changes using structural ensembles. *Nat Methods*, **6**, 3–4.

[318] Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0. *Bioinformatics*, **25**, 2537–2543.

[319] Potapov, V., Cohen, M., Inbar, Y., and Schreiber, G. (2010) Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. *BMC Bioinformatics*, **11**, 374.

[320] Seeliger, D. and de Groot, B. L. (2010) Protein thermostability calculations using alchemical free energy simulations. *Biophys J*, **98**, 2309–2316.

[321] Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, **79**, 830–838.

[322] Wickstrom, L., Gallicchio, E., and Levy, R. M. (2012) The linear interaction energy method for the prediction of protein stability changes upon mutation. *Proteins*, **80**, 111–125.

[323] Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014) mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.

[324] Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014) Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*, **42**, W314–W319.

[325] Giollo, M., Martin, A. J. M., Walsh, I., Ferrari, C., and Tosatto, S. C. E. (2014) Neemo: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics*, **15 Suppl 4**, S7.

[326] Frappier, V. and Najmanovich, R. J. (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Comput Biol*, **10**, e1003569.

[327] Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., and Lackner, P. (2015) Maestro–multi agent stability prediction upon point mutations. *BMC Bioinformatics*, **16**, 116.

[328] Stehman, S. V. (1997) Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62**, 77–89.

[329] Myers, J. L., Well, A. D., and Lorch, R. F. (2010) *Research Design and Statistical Analysis*. Routledge.

[330] Wan, J., Kang, S., Tang, C., Yan, J., Ren, Y., Liu, J., Gao, X., Banerjee, A., Ellis, L. B. M., and Li, T. (2008) Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res*, **36**, e22.

[331] Yang, J., Roy, A., and Zhang, Y. (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

[332] Qin, S. and Zhou, H.-X. (2007) meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, **23**, 3386–3387.

[333] Ishida, T. and Kinoshita, K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–1348.

[334] Kozlowski, L. P. and Bujnicki, J. M. (2012) Metadisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 111.

[335] Emily, M., Talvas, A., and Delamarche, C. (2013) Metamyl: a meta-predictor for amyloid proteins. *PLoS One*, **8**, e79722.

[336] Foit, L., Morgan, G. J., Kern, M. J., Steimer, L. R., von Hacht, A. A., Titchmarsh, J., Warriner, S. L., Radford, S. E., and Bardwell, J. C. A. (2009) Optimizing protein stability in vivo. *Mol Cell*, **36**, 861–871.

[337] Deng, Z., Huang, W., Bakkalbasi, E., Brown, N. G., Adamski, C. J., Rice, K., Muzny, D., Gibbs, R. A., and Palzkill, T. (2012) Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. *J Mol Biol*, **424**, 150–167.

[338] Wei, Q. and Dunbrack, R. L. (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, **8**, e67863.

[339] Borgo, B. and Havranek, J. J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. *Proc Natl Acad Sci U S A*, **109**, 1494–1499.

[340] Gapsys, V., Michielssens, S., Seeliger, D., and de Groot, B. L. (2015) pmx: Automated protein structure and topology generation for alchemical perturbations. *J Comput Chem*, **36**, 348–354.

[341] Song, X., Wang, Y., Shu, Z., Hong, J., Li, T., and Yao, L. (2013) Engineering a more thermostable blue light photo receptor bacillus subtilis ytva lov domain by a computer aided rational design method. *PLoS Comput Biol*, **9**, e1003129.

[342] Vihinen, M. (2014) Majority vote and other problems when using computational tools. *Hum Mutat*, **35**, 912–914.

[343] Rees, D. C. and Robertson, A. D. (2001) Some thermodynamic implications for the thermostability of proteins. *Protein Sci*, **10**, 1187–1194.

[344] Allen, B. D., Nisthal, A., and Mayo, S. L. (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci U S A*, **107**, 19838–19843.

[345] Davey, J. A. and Chica, R. A. (2014) Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins*, **82**, 771–784.

[346] Christensen, N. J. and Kepp, K. P. (2012) Accurate stabilities of laccase mutants predicted with a modified foldx protocol. *J Chem Inf Model*, **52**, 3028–3042.

[347] Wimley, W. C., Creamer, T. P., and White, S. H. (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*, **35**, 5109–5124.

[348] Darby, N. and Creighton, T. (1993) *Protein Structure*. IRL Press at Oxford University Press.

[349] Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

[350] Kim, D. E., Fisher, C., and Baker, D. (2000) A breakdown of symmetry in the folding transition state of protein l. *J Mol Biol*, **298**, 971–984.

[351] Chen, J. and Stites, W. E. (2001) Higher-order packing interactions in triple and quadruple mutants of staphylococcal nuclease. *Biochemistry*, **40**, 14012–14019.

[352] Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., and Ventura, S. (2015) Aggrescan3d (a3d): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res*, **43**, W306–W313.

[353] Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015) The camsol method of rational design of protein mutants with enhanced solubility. *J Mol Biol*, **427**, 478–490.

[354] Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014) Pasta 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res*, **42**, W301–W307.

[355] Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, **22**, 1302–1306.

[356] Goldschmidt, L., Teng, P. K., Riek, R., and Eisenberg, D. (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc Natl Acad Sci U S A*, **107**, 3487–3492.

[357] Tartaglia, G. G. and Vendruscolo, M. (2008) The zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev*, **37**, 1395–1401.

[358] Gilis, D., McLennan, H. R., Dehouck, Y., Cabrita, L. D., Rooman, M., and Bottomley, S. P. (2003) In vitro and in silico design of alpha1-antitrypsin mutants with different conformational stabilities. *J Mol Biol*, **325**, 581–589.

[359] Cabrita, L. D., Gilis, D., Robertson, A. L., Dehouck, Y., Rooman, M., and Bottomley, S. P. (2007) Enhancing the stability and solubility of tev protease using in silico design. *Protein Sci*, **16**, 2360–2367.

[360] Yang, D.-F., Wei, Y.-T., and Huang, R.-B. (2007) Computer-aided design of the stability of pyruvate formate-lyase from escherichia coli by site-directed mutagenesis. *Biosci Biotechnol Biochem*, **71**, 746–753.

[361] Zhang, S.-B. and Wu, Z.-L. (2011) Identification of amino acid residues responsible for increased thermostability of feruloyl esterase a from aspergillus niger using the popmusic algorithm. *Bioresour Technol*, **102**, 2093–2096.

[362] Komor, R. S., Romero, P. A., Xie, C. B., and Arnold, F. H. (2012) Highly thermostable fungal cellobiohydrolase i (cel7a) engineered using predictive methods. *Protein Eng Des Sel*, **25**, 827–833.

[363] Silva, I. R., Larsen, D. M., Jers, C., Derkx, P., Meyer, A. S., and Mikkelsen, J. D. (2013) Enhancing rgi lyase thermostability by targeted single point mutations. *Appl Microbiol Biotechnol*, **97**, 9727–9735.

[364] Wijma, H. J., Floor, R. J., Jekel, P. A., Baker, D., Marrink, S. J., and Janssen, D. B. (2014) Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel*, **27**, 49–58.

[365] Deng, Z., Yang, H., Li, J., Shin, H.-D., Du, G., Liu, L., and Chen, J. (2014) Structure-based engineering of alkaline alpha-amylase from alkaliphilic alkalimonas amylolytica for improved thermostability. *Appl Microbiol Biotechnol*, **98**, 3997–4007.

[366] Larsen, D. M., Nyffenegger, C., Swiniarska, M. M., Thygesen, A., Strube, M. L., Meyer, A. S., and Mikkelsen, J. D. (2015) Thermostability enhancement of an endo-1,4-beta-galactanase from talaromyces stipitatus by site-directed mutagenesis. *Appl Microbiol Biotechnol*, **99**, 4245–4253.

[367] Heselpoth, R. D., Yin, Y., Moult, J., and Nelson, D. C. (2015) Increasing the stability of the bacteriophage endolysin plyc using rationale-based foldx computational modeling. *Protein Eng Des Sel*, **28**, 85–92.

[368] Verlet, L. (1967) Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, **159**, 98–103.

[369] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics*, **23**, 327–341.

[370] Darden, T., York, D., and Pedersen, L. (1993) Particle mesh ewald: An n*log(n) method for ewald sums in large systems. *J. Chem. Phys.*, **98**, 10089–10092.

[371] Groenhof, G. (2013) Introduction to qm/mm simulations. *Methods Mol Biol*, **924**, 43–66.

[372] Steinbrecher, T. and Elstner, M. (2013) Qm and qm/mm simulations of proteins. *Methods Mol Biol*, **924**, 91–124.

[373] Karplus, M. and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, **9**, 646–652.

[374] Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys*, **41**, 429–452.

[375] Pikkemaat, M. G., Linssen, A. B. M., Berendsen, H. J. C., and Janssen, D. B. (2002) Molecular dynamics simulations as a tool for improving protein stability. *Protein Eng*, **15**, 185–192.

[376] Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011) How fast-folding proteins fold. *Science*, **334**, 517–520.

[377] Pan, A. C., Weinreich, T. M., Piana, S., and Shaw, D. E. (2016) Demonstrating an order-of-magnitude sampling enhancement in molecular dynamics simulations of complex protein systems. *J Chem Theory Comput*, **12**, 1360–1367.

[378] Torrie, G. and Valleau, J. (1977) Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, **23**, 187–199.

[379] Beutler, T. C. and van Gunsteren, W. F. (1994) The computation of a potential of mean force: Choice of the biasing potential in the umbrella sampling technique. *J. Chem. Phys.*, **100**, 1492.

[380] Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., and Rosenberg, J. M. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comp. Chem.*, **13**, 1011–1021.

[381] Bussi, G., Laio, A., and Parrinello, M. (2006) Equilibrium free energies from nonequilibrium metadynamics. *Phys Rev Lett*, **96**, 090601.

[382] Laio, A. and Gervasio, F. (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.*, **71**, 126601.

[383] Barducci, A., Bussi, G., and Parrinello, M. (2008) Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys Rev Lett*, **100**, 020603.

[384] Comer, J., Gumbart, J. C., Hénin, J., Lelièvre, T., Pohorille, A., and Chipot, C. (2015) The adaptive biasing force method: everything you always wanted to know but were afraid to ask. *J Phys Chem B*, **119**, 1129–1151.

[385] Bagal, D., Kitova, E. N., Liu, L., El-Hawiet, A., Schnier, P. D., and Klassen, J. S. (2009) Gas phase stabilization of noncovalent protein complexes formed by electrospray ionization. *Anal Chem*, **81**, 7801–7806.

[386] Dror, R. O., Young, C., and Shaw, D. E. (2011) *"Anton, a Special-Purpose Molecular Simulation Machine," Encyclopedia of Parallel Computing*. Spinger.

[387] Liu, L., Michelsen, K., Kitova, E. N., Schnier, P. D., and Klassen, J. S. (2012) Energetics of lipid binding in a hydrophobic protein cavity. *J Am Chem Soc*, **134**, 3054–3060.

[388] Kitova, E. N., Seo, M., Roy, P.-N., and Klassen, J. S. (2008) Elucidating the intermolecular interactions within a desolvated protein-ligand complex. an experimental and computational study. *J Am Chem Soc*, **130**, 1214–1226.

[389] Jalili, N. (2014) *Computational Study of Bovine beta-Lactoglobulin Complexes with Fatty Acids*. Master's thesis, University of Alberta.

[390] van Zon, R. and Schofield, J. (2010) Constructing smooth potentials of mean force, radial distribution functions, and probability densities from sampled data. *J Chem Phys*, **132**, 154110.

[391] Tritzant-Martinez, Y., Zeng, T., Broom, A., Meiering, E., Roy, R. J. L., and Roy, P.-N. (2013) On the analytical representation of free energy profiles with a morse/long-range model: application to the water dimer. *J Chem Phys*, **138**, 234103.

[392] Resh, M. D. (2006) Trafficking and signaling by fatty-acylated and prenylated proteins. *Nat Chem Biol*, **2**, 584–590.

[393] Fan, Z., Dror, R. O., Mildorf, T. J., Piana, S., and Shaw, D. E. (2015) Identifying localized changes in large systems: Change-point detection for biomolecular simulations. *Proc Natl Acad Sci U S A*, **112**, 7454–7459.

[394] Holm, L. and Rosenström, P. (2010) Dali server: conservation mapping in 3d. *Nucleic Acids Res*, **38**, W545–W549.

[395] Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F.-Y. (2011) R.e.d. server: a web service for deriving resp and esp charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res*, **39**, W511–W517.

[396] Dupradeau, F.-Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., and Cieplak, P. (2010) The r.e.d. tools: advances in resp and esp charge derivation and force field library building. *Phys Chem Chem Phys*, **12**, 7821–7839.

[397] Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*, **25**, 247–260.

[398] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general amber force field. *J Comput Chem*, **25**, 1157–1174.

[399] Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., and Woods, R. J. (2008) Glycam06: a generalizable biomolecular force field. carbohydrates. *J Comput Chem*, **29**, 622–655.

[400] Eastman, P., et al. (2013) Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *J Chem Theory Comput*, **9**, 461–469.

[401] Krishna, M. M. G. and Englander, S. W. (2005) The n-terminal to c-terminal motif in protein folding and function. *Proc Natl Acad Sci U S A*, **102**, 1053–1058.

[402] Davis, I. W., Arendall, W. B., Richardson, D. C., and Richardson, J. S. (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*, **14**, 265–274.

[403] Smith, C. A. and Kortemme, T. (2011) Predicting the tolerated sequences for proteins and protein interfaces using rosettabackrub flexible backbone design. *PLoS One*, **6**, e20451.

[404] Chitsaz, M. and Mayo, S. L. (2013) Grid: a high-resolution protein structure refinement algorithm. *J Comput Chem*, **34**, 445–450.