Iterative Multistate Negative Design of protein folds

Javier Castellanos

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

David Baker, Chair

Barry Stoddard

Wim Hol

Program Authorized to Offer Degree:

Biochemistry

University of Washington

Abstract

Iterative Multistate Negative Design of protein folds

Javier Castellanos

Chair of Supervisory Committee:

Professor David Baker

Department of Biochemistry

One of the challenges of computational de novo protein design is to find a protein sequence that adopts a desired conformation disfavoring alternative low energy conformations. A protein adopts a specific fold when the interactions that stabilize the native state collectively outweigh the interactions present in all the possible alternative states. To stabilize a protein in a specific fold one can either reduce the energy of the protein in the target conformation(positive design) or increase the energy of the alternative conformations(negative design). Here we describe a new protocol that simultaneously stabilizes the target state and destabilizes the possible alternative low energy states of de novo design proteins. The new protocol performs a contact analysis of the alternative states

to identify residues that are participating undesired contacts and then performs a multi state sequence optimization of those positions. After iterations of sampling,analysis and optimization the protocol is able to increase the energy of the alternative states and improve the sampling close to de novo design structure.

TABLE OF CONTENTS

Acknowledgments

I would like to express my sincere thanks to my advisor, David Baker. He welcomed me into his lab were I had the privilege of seeing how cutting-edge science is made. His enthusiasm and optimism for the field of protein design ca be contagious and inspiring. I would also like to thank the University of Washington and specially the Biochemistry department for providing such a wonderful learning environment. I am grateful for the insights to academic life that I got from many the post docs from the Baker lab, they helped me to discover what I wanted to do in the next chapter of my life. My parents and brothers deserve a lot of credit, they have supported me tremendously in my quest to satisfy my intellectual curiosity. Finally aI would like to thank Katie, my wife,who has supported me throughout graduate school, always giving me a hug when need it.

Chapter One: Introduction to Computational *de novo* protein design.

Significant progress has been made in the field of protein design during the last three decades. The problem of protein design is to find amino acid sequences that are compatible with a given protein structure[1]. Early work done by Ponder and Richards[2] introduced the basic elements of modern protein design. Their strategy consisted of the use of a side-chain rotamer library and an energy function to sample the allowed sequences of a particular structure. Hellinga and Richards refined this strategy by including simulated annealing as the optimization method that guides the design process.[3] The new method consisted of random walks in sequence space generated by movements either in rotamer space (modification of the rotamer at a random position) or vibrational space (small changes in the side chain torsion angles) that are selected using the Metropolis-Hastings algorithm.

Most of the early work on protein design was focused on the redesign of the hydrophobic core of stable proteins[4, 5] and helical bundles[6-8]. Desjarlais and Handel redesigned variants of the core of the page 434 cro protein, a five helix DNA-binding protein. Five, seven and eight mutations were made to the core of the cro protein and two of the designs had similar stabilities when compared to the native cro protein[4]. Lazar and Handel use the same strategy to redesign

the core of ubiquitin, and all the designs were destabilized compared to native structure. This suggested that β-sheets required more stringent packing of the side chains than α-helices[5]. Bryson and DeGrado designed a novel globular protein by redesigning a coiled-coil peptide into a three-helix bundle[6]. Harbury and Kim designed novel right-handed coiled-coils from parametrically generated α-helices.[7] This work demonstrated that it was possible to design new proteins from scratch using computer-generated protein backbones.

The Mayo group developed an alternative strategy for protein design based on the Dead End Elimination Theorem (DEE)[9, 10]. DEE is used to shrink the size of conformational space by reducing the number of available rotamers. Using this technique Dahiyat and Mayo designed a completely new sequence for a ββα protein based on the zinc finger domain[11]. This design can be considered as the first *de novo* designed protein since its sequence contained no information from the natural evolution of the fold.

One of the landmarks of the protein design field was the design of Top7, the first artificial protein to adopt a novel fold not observed in nature[12]. Kuhlman and Baker sketched out a topology for an α/β protein that was not present in the Topology of Protein Structure Server (TOPS)[13]. The protein backbone for Top7 was assembled from fragments of the proteins present in the Protein Data Bank following a strategy that resembled the protein structure prediction protocol developed by the Baker Group[14]. The sequence was designed using the

RosettaDesign protocol that uses a Metropolis Monte Carlo procedure[15]. The Monte Carlo does not guarantee convergence to a globally optimum minimum but has the advantage of being very fast and that it can be trivially parallelized. Through the designing of a novel topology the authors found an excellent way to test the maturity of the techniques and tools of protein design

One of the most interesting applications of computational protein design is to create new or improved enzymes for therapeutical and industrial purposes. Bolon and Mayo used the DEE optimization strategy to add histidine-mediated nucleophilic hydrolysis of p-nitrophenyl acetate catalytic activity to the inert thioredoxin protein[16]. Kaplan and DeGrado designed *de novo* diiron proteins able to catalyze a phenol oxidase reaction[17]. Zanghellini and Baker developed a geometric hashing algorithm that allowed rapid identification of proteins in the PDB that might accommodate the active of an arbitrary reaction[18]. The Baker group successfully designed enzymes that catalyzed retro-aldol[19], Kemp elimination[20] and Diels-Alder reaction[21] using this method.

Design of protein-protein interfaces is another interesting area of protein design because of it applicability for therapeutics and material design. Kortemme and Baker developed computational methods to predict energy ‘hot spots’ in the protein-protein interface[22] by performing a computational alanine scan. Shifman and Mayo increased the affinity of calmodulin to one of its targets by redesigning the interface using a physically based force field and the DEE

theorem[23]. Huang and Mayo created an artificial heterodimer by designing an interface between two monomeric proteins[24]. Most recently the Baker group has one-component[25] and two-component[26] molecular cages by designing interfaces between cyclic homooligomers.

Most of the advances of the protein design field have been focused on the redesign of existing proteins. *De novo* protein design has progressed more slowly because it has the additional difficulty of generating feasible protein backbones that lack any evolutionary sequence information. The advantage of *de novo* designed proteins with custom made backbones is that they can be specifically tailored for the desired function. Progress in *de novo* design will open the door for a new generation of enzymes and protein-based materials by allowing to do more than just repurposing natural proteins.

The Challenge of de novo design

Computational *de novo* protein design is a stringent test of our knowledge of the principles that govern structural biology. Full understanding of these principles will allow us one day to design custom made proteins for industrial and therapeutic purposes. The design of *de novo* proteins consists of three steps: backbone generation, sequence optimization and *in silico* validation[8, 12, 27].

Backbone generation

Backbone generation is the process of building protein backbones for a particular target topology. Backbones can be generated either parametrically[7, 28] or through combinatorial sampling of fragments from known structures[12]. In Rosetta[29], protein backbones are generated through combinatorial sampling of fragments of proteins from the PDB whose secondary structure matches the one of the target topologies. The fragments are randomly inserted into an extended chain using an algorithm that resembles the Rosetta *de novo* protein structure prediction protocol[12, 27]. The length of the secondary structure elements need to be optimized in order to generate protein backbones consistently. The length of the loop regions can be derived using the rules defined by Koga, Tatsumi-Koga and Baker (Figure 1)[27]. These rules predict the orientation of secondary structure elements based on the lengths of the loops that connects them. Following these rules increases the probability of sampling the desired conformation during the Monte Carlo simulation. Once backbones with the desired topology have been produced, the next step is to find sequences that are compatible with the backbone geometry.

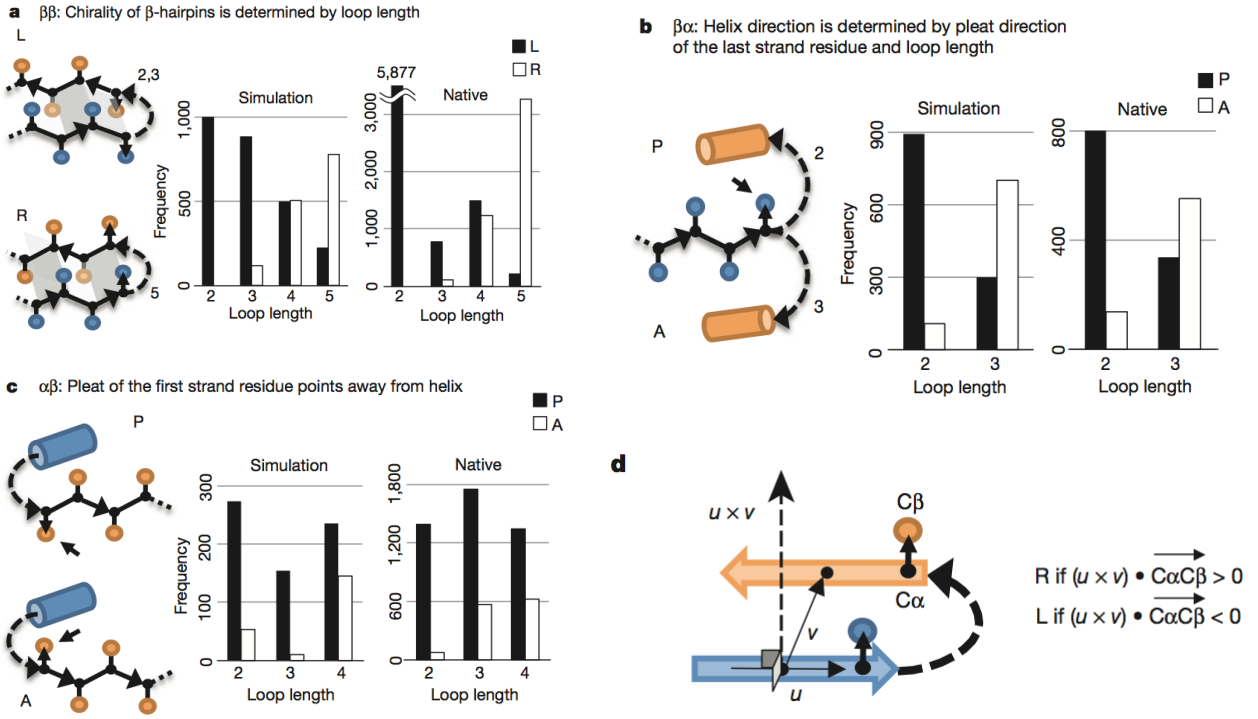


Figure 1. Koga and Baker fundamental rules. a, **ββ**-rule. L (left-handed) and R (right-handed) **ββ**-units are illustrated (see Fig. 1d for chirality definition). The dependence of chirality on loop length is shown in the histograms. b, **βα**-rule. P (parallel) and A (antiparallel) **βα**-units are illustrated. The dependence of orientation (P versus A) on loop length is shown in the histograms. c, ab-rule. d, Chirality (L versus R) of a **ββ**-unit. The chirality is defined on the basis of the orientation of the Cα-to-C**β** vector, CαC**β**, of the strand residue preceding or following the connecting loop. u is a vector along the first strand and v is a vector from the centre of the first strand to the centre of the second strand. Figure adapted from Koga et al. *Principles for designing ideal protein structures*(2013).

Sequence Design

Protein sequence design is an optimization problem where the amino acid sequence of a protein is optimized to adopt a target conformation. Monte Carlo[30-32] and Dead-End elimination theorem[9, 10, 33-37] are the most widely used methods in sequence design. The optimization process is guided by a score function that can be either a molecular mechanics potential energy function,a knowledge-based potential[38-40] or a hybrid between both approaches[15].

Molecular mechanics potential energy functions incorporate ‘bonded’ and ‘non-bonded’ terms that approximate the energetic state of the system[41]. Bonded terms apply to sets of two or four atoms that are covalently linked and constraint the bond lengths and angles near their equilibrium values. The bonded terms also include a torsion potential that models the periodic energy barriers encountered during bond rotation. The non-bonded terms consist of the Lennard–Jones (LJ) function (which includes van der Waals attraction and repulsion) and Coulomb’s law[42, 43].

Knowledge-based (or statistical) potentials take advantage of the vast amount of data that has been collected in the Protein Data Bank. These potentials use the Boltzmann equation, $\Delta G = -RT\ln(p_{obs}/p_{exp})$, to derive effective potentials from statistics of observables in a structural database[39, 44].

Hybrid score functions, like the Rosetta score function, are a combination of empirical terms from molecular mechanics potential energy functions and knowledge-based potentials[11, 15, 39]. The advantage of this type of score function is that it can correct for errors in the energy function by incorporating statistical data, and can be tuned for specific problems such as calculating binding energies or sequence design.

The Rosetta score function is a linear combination that includes, among other terms, a standard 12–6 Lennard–Jones potential with van der Waals radii and well depths from the CHARMM19 parameter set, a repulsive term that connects with the 12–6 potential at $E = 0$, a solvation term computed using the Lazaridis–Karplus implicit solvation model, an approximation to electrostatic interactions in proteins based on PDB statistics[15], and a orientation-dependent hydrogen bonding potential[45].

Two strategies used in protein sequence design are fixed and flexible backbone design. In fixed-backbone design the atoms of the protein backbone are held invariant and only the side chain atoms are allowed to move during the optimization process[2, 3, 46]. Flexible-backbone design is usually performed by reducing the repulsive term of the Lennard-Jones potential[11] or through cycles of fix-backbone design followed by full-atom gradient based minimization[12, 27].

Fixed-backbone design in Rosetta is performed using Monte Carlo optimization and simulated annealing[15]. In each step of the optimization process, the side chain of a random position is replaced by other rotamer obtained from the Dunbrack rotamer library[47] and the resulting structure is accepted or rejected according to the Metropolis criterion[48]. After a fixed number of steps, a side-chain only torsional conjugate-gradient step minimization is performed to reduce the repulsion between side chains that are packing against each other[15].

In Silico Validation

Computationally designed sequences can be tested *in silico* using a protein structure prediction protocol. Structure prediction is the problem of predicting the three dimensional structure of protein from its amino acid sequence. Structure prediction is the inverse problem of protein design, and as such it can be used to test if designed proteins are predicted to adopt the target fold.

The two main approaches in protein structure prediction are homology modeling and *ab initio* prediction. In homology modeling a multiple sequence alignment is used for the selection of templates that are used to construct the model[49-51]. This strategy is based on the observation done by Chothia and Lesk that structural similarity is more conserved that sequence similarity[52]. *Ab initio* structure prediction methods are based on the thermodynamic hypothesis formulated by Anfinsen[53] that stipulates that the native structure corresponds

to the global free energy minimum under the given set of conditions. These methods use a sampling method to explore conformational space, and a score function to approximate the free energy of the conformation[54, 55].

Ab initio structure prediction is a stringent method to test the energetic consistency of the design. Sampling near the designed structure can be disfavored if the designed protein has destabilizing interactions (Figure 2. A). It is also possible that the designed sequence adopts an alternative conformation that has lower energy (Figure 2.B). A well designed protein sequence should have a funnel-like energetic landscape with the designed conformation occupying the bottom of the energy funnel (Figure 2.C). The next chapter describes the use of methods described in this chapter for the design of a *de novo* protein that adopts the P-loop topology.

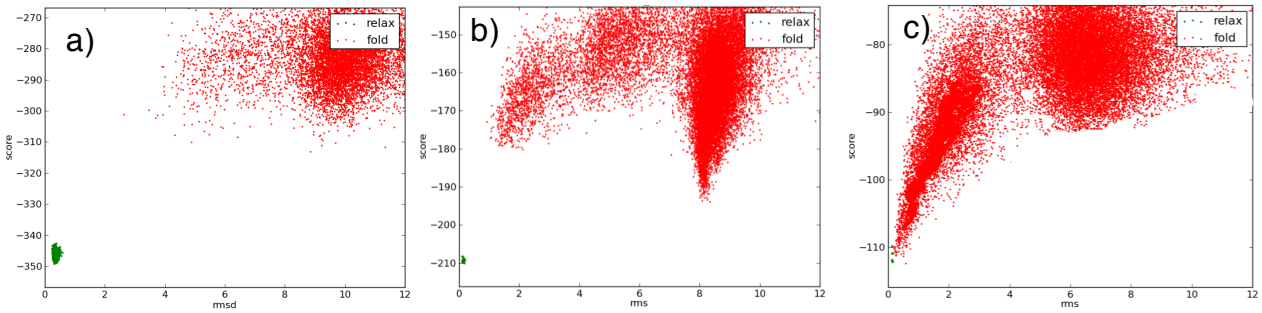


Figure 2. Example of three different energy landscapes. Each red point in the plot represents a structure generated by the Rosetta *ab initio* structure prediction protocol, the green points are trajectories of the designed structure after optimization with conjugated gradients minimization. The x-axis is the *root mean square deviation* (RMSD) of the structure when compared to the conformation of the designed structure and the y-axis is the score of the structure according to Rosetta's score function. a) Energy landscape of a designed protein with poor sampling in the near-target region, most of the sampling is concentrated at 10Å RMSD. b) Energy landscape of a designed protein with and alternative minima at 8Å RMSD. c) Energy landscape of a designed protein with a sharp energy funnel close to the designed structure.

Chapter Two: *de novo* design of the P-loop 2x3 fold

The P-loop 2x3 is a Rossmann-like fold composed of a core of five parallel beta strands and five alpha helices (Figure 3). This topology is mostly observed in phosphate binding proteins like kinases and nucleoside binding proteins. The P-loop fold has a competing topology, the Flavodoxin fold, that needs to be destabilized to ensure proper folding to the target topology. The Flavodoxin fold and the P-loop 2x3 have a similar topology. The main difference between these folds is that strands one and three are swapping their position in the beta sheet (Figure 4). Flavodoxins are electron-transfer proteins involved in a variety of photosynthetic and non-photosynthetic reactions in procaryotes.These proteins obtain their redox activity through binding of flavin mono nucleotide cofactor[64].

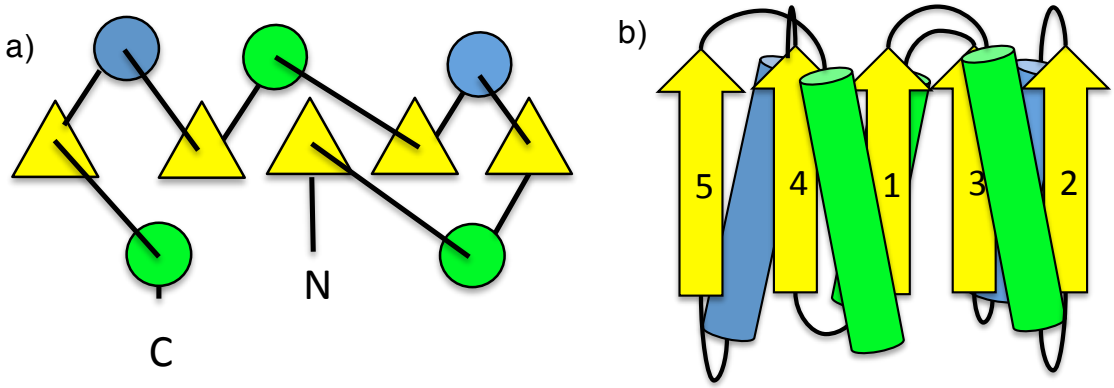


Figure 3. P-loop 2x3 topology. a)Diagram of the top view of the P-loop 2x3 fold, alpha helices are shown in circles and beta strands in triangles. The helices in blue are forming contiguous *beta-alpha-beta* domains while the helices in blue represent *crossover* domains where the beta strands are not pairing with each other. b) Side view diagram of the P-loop 2x3 fold. The numeration in the beta strands represents the relative position of the strand to the N-terminus.

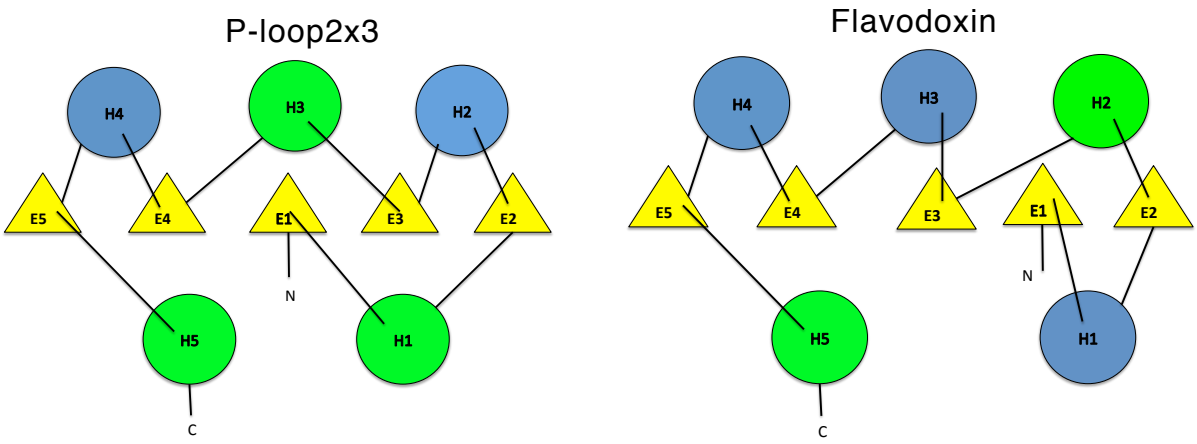


Figure 4. Comparison between the P-loop 2x3 topology(left) and the Flavodoxin (right) topology. The Flavodoxin topology differs from the P-loop2x3 topology in that strands one and three are swapped in their relative position in the beta sheet.

Computational Design

The first step of the design process consisted of optimizing the length of the secondary structure elements. This was done by running several folding simulations that differ on length of the secondary structure elements that composed the structure. The combination of elements that generated the P-loop 2x3 topology, while disfavoring the competitor structure, was selected. Since the P-loop 2x3 topology has many strands that can pair in multiple orientations it was necessary to decompose the optimization process into three different steps in order to increase the sampling efficiency.

The first step consisted of sampling from strand one to strand three, the second used the best combination of step one and added helix three and strand four and the third step used the best combination of step two and added helix four and strand five (Figure 5).

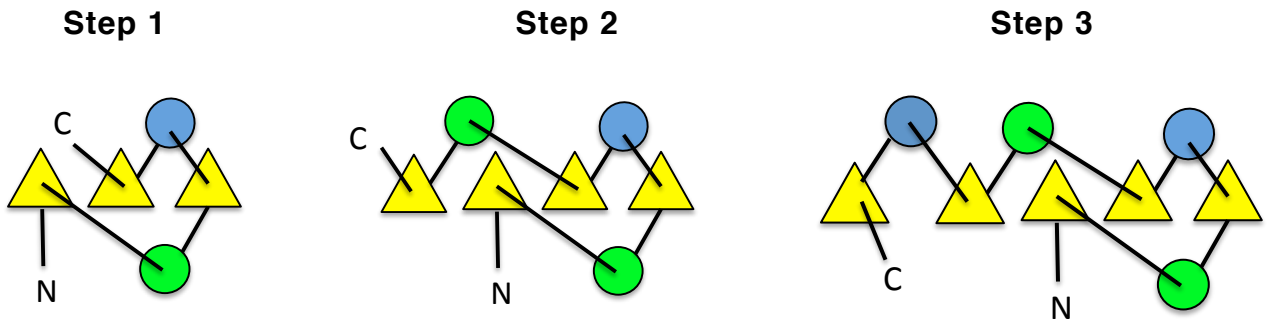


Figure 5. Steps used for the combinatorial sampling of the secondary structure lengths.

The structures generated by the folding simulation for each combination where analyzed by counting the observations of the target fold and the competing fold. For step one (Figure 6) the sampled combinations varied the length of the first helix and the adjacent loops. The strands were fixed at lengths 7,5, and 7 to disfavor the swapping of strands one and three because that would necessarily reduce the number of backbone-backbone hydrogen bonds. The combination that was selected for this step had the highest ratio of samples in the target conformation to competing the conformation. The same procedure was repeated for steps two and three obtaining the length combinations shown in Figure 7.

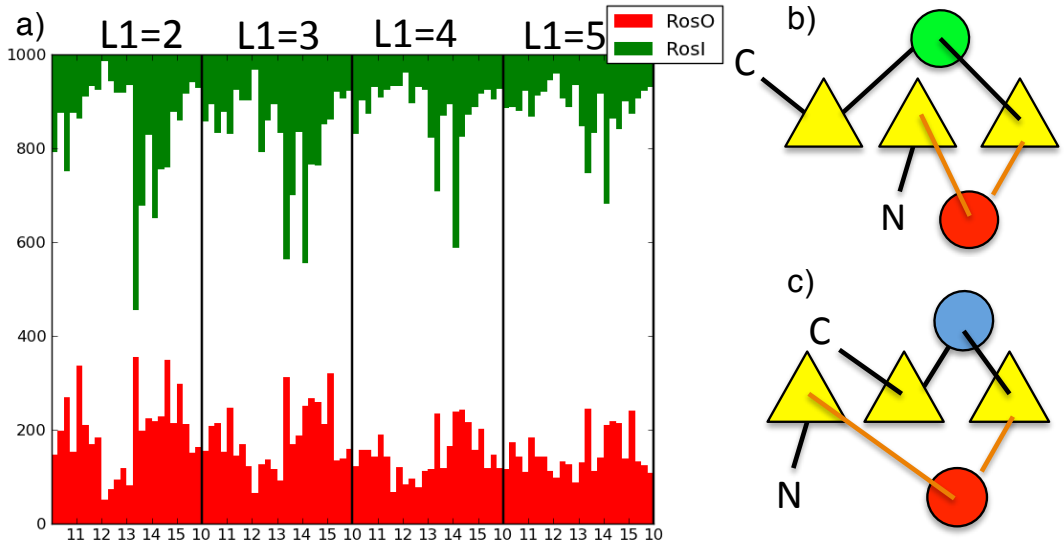


Figure 6. Step 1 of the secondary structure length optimization. Folding simulations were ran for combinations that sampled for the optimal length of the first helix (*6.c*, red circle) and the adjacent loop elements (*6.c,* orange lines). The N-termini and C-termini loops of helix one were sampled for lengths between 2 to 5 residues and the helix was sampled between 10 and 15 residues producing a total of 96 possible combinations. 8000 folding simulations were run for each of the 96 combinations and the results are shown aggregated in figure 6.a. The major line divisions in figure 6.a represent the length of the N-termini loop connected to the first helix (label on the top). Each major division is further divided into 6 bins labeled from 10 to 15(*6.a* bottom) that represent the length of the first helix and each of this sub-bins is further divided in 4 bin(unlabeled) that represent the length of the loop connected to the C-termini of the first helix. The red bars represent the number of counts for the desired topology (*6.c*) and the green bars the number of counts of the alternative topology (*6.b*).

The optimized length combination of secondary structure elements was used to generate of a set of protein backbones. The backbones were used for the design of several thousand sequences using the flexible-backbone design protocol. The designed models were filtered by total score and RosettaHoles score[56]. The top 20 designs according to this metrics were validated *in silico* using the *ab initio* structure prediction protocol on the Rosetta@HOME distributed computing platform. Most of the tested designs didn't produce a clear energy funnel but a

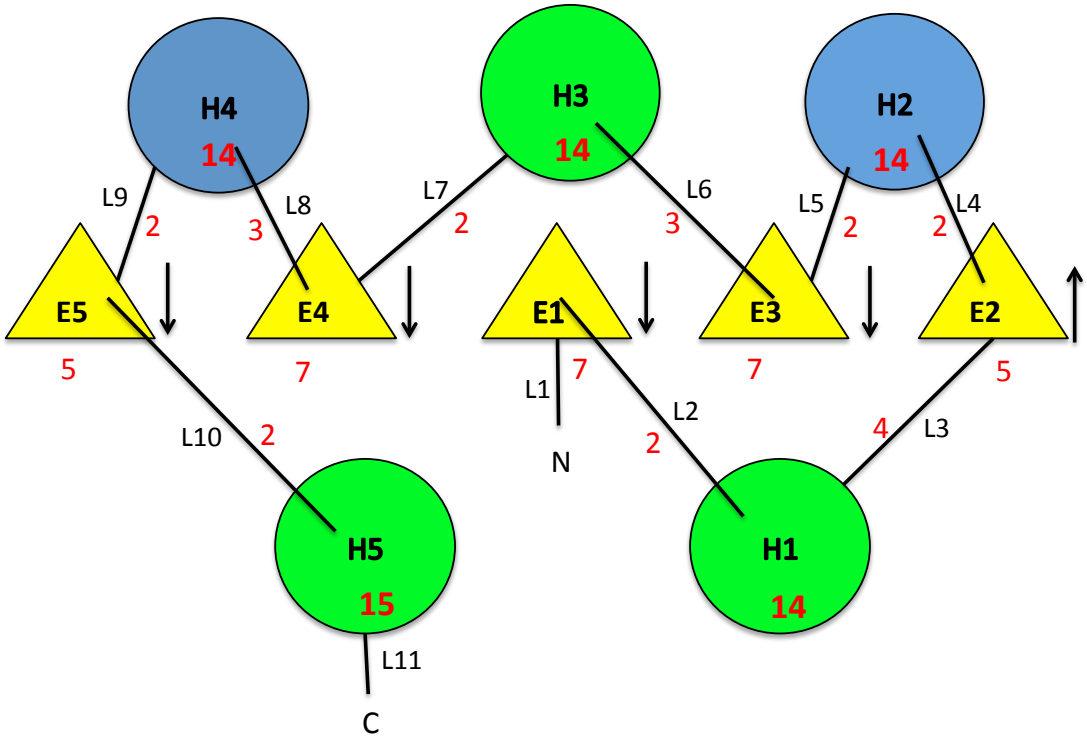


Figure 7. Top view diagram of the P-loop 2x3 topology with the optimized secondary structure element lengths. Helices are shown in circles and beta strands with triangles. The arrows in the left of the triangles indicate the direction of the pleat of the last strand residue. In red are the number of residues of the different secondary structure elements. The length of all the loops are consistent with the Koga and Baker fundamental rules.

few of them looked promising. Those designs were manually redesigned over 70 design iterations of the cycle shown in figure 8. Twelve designs where picked for experimental testing based on the criteria that they displayed consistent energy funnel and that the desired topology was being sampled at lower energy than the alternative topology (Figure 9).

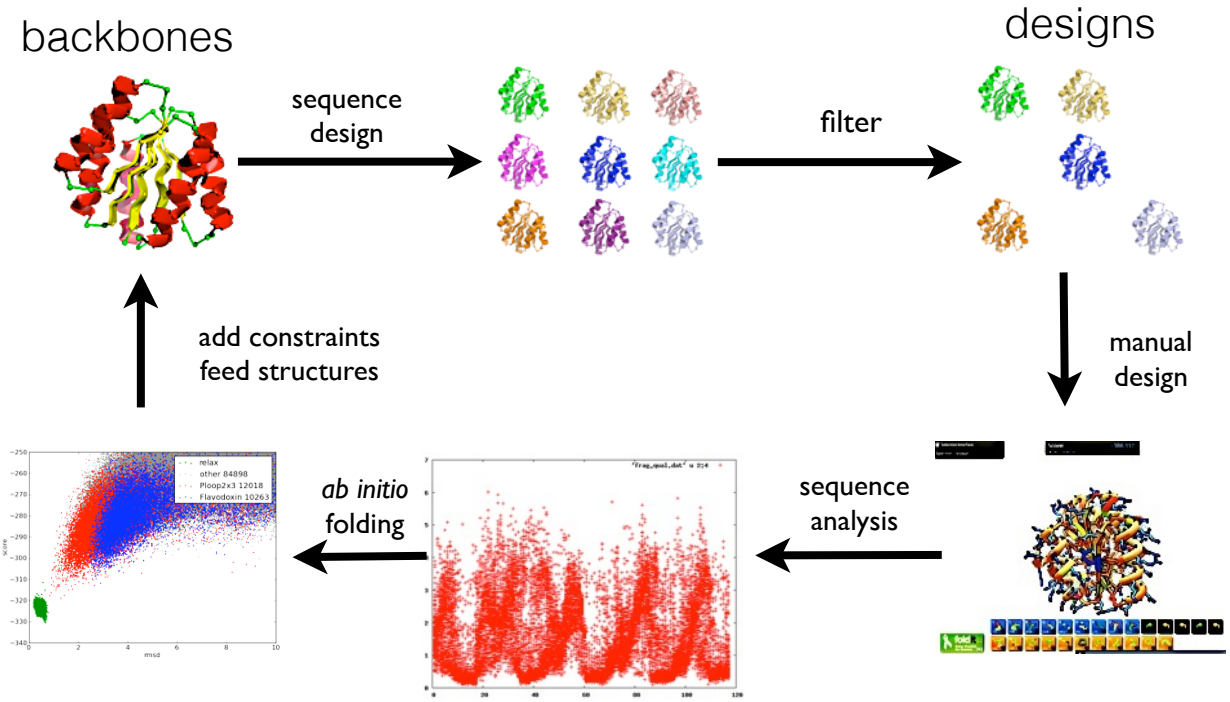


Figure 8. *de novo* protein design workflow. Protein backbones are generated using the method described in chapter I (*top left*).These backbones are then used for sequence design and posterior filtering. The most promising designs then go through a round of manual redesign using the stand alone *Foldit* application. The new sequences are then further analyzed for their compatibility with fragments on the PDB and then validated using the *ab initio* structure prediction protocol. If the design is successful the predicted backbone is reused in a new iteration of the design protocol. The points shown in green in the energy landscape shown in this figure(*bottom left)* represent minimization trajectories of the design model, the red and blue points represent the desired topology and the alternative topology respectively.

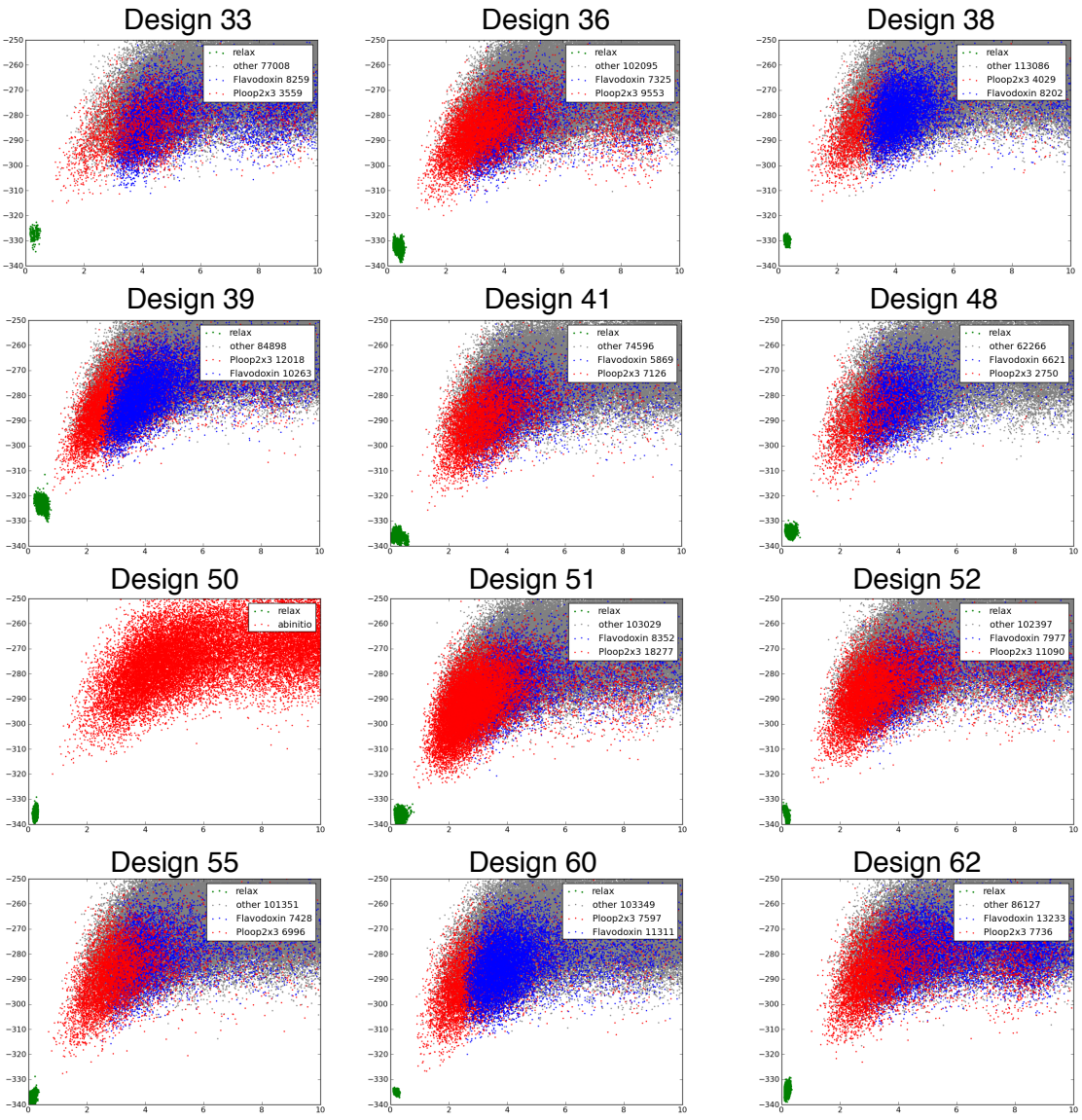


Figure 9 Energy landscape of the 12 designs selected for experimental validation. The structures colored in red are adopting the P-loop 2x3 fold the ones in blue adopt the competing fold. The x-axis show the rmsd of the models when compared to the design structure and the y-axis the energy of the model as calculated by the Rosetta score function.

Experimental Results

The twelve that were picked for experimental validation were expressed, purified and characterized. Circular dichroism (Figure 11) and spectroscopy and size-exclusion chromatography combined with multi-angle light scattering (SEC-MALS) was used for the characterization. Seven of the designs expressed where soluble enough for experimental characterization. The results are shown in Figure 10.

design	expressed	solubility	αβ-protein circular dichroism spectrum	Oligomerization
33	no	n/a	n/a	n/a
36	yes	high	no	dimer
38	no	n/a	n/a	n/a
39	yes	low	no	dimer
41	no	n/a	n/a	n/a
48	yes	high	yes	monomer/dimer
50	yes	high	yes	monomer
51	no	n/a	n/a	n/a
52	yes	high	yes	monomer
55	yes	low	yes	aggregate
60	yes	low	yes	aggregate
62	no	n/a	n/a	n/a

Figure 10. Aggregated results of the experimental characterization of the 12 ordered P-loop2x3 designs. Expression of the protein was determined through SDS-PAGE, solubility was classified as high if concentration after the elution from the Nickel affinity column with 12 mL of elution buffer was higher than 2 mg/mL. The oligomerization state was determined by SEC-MALS.

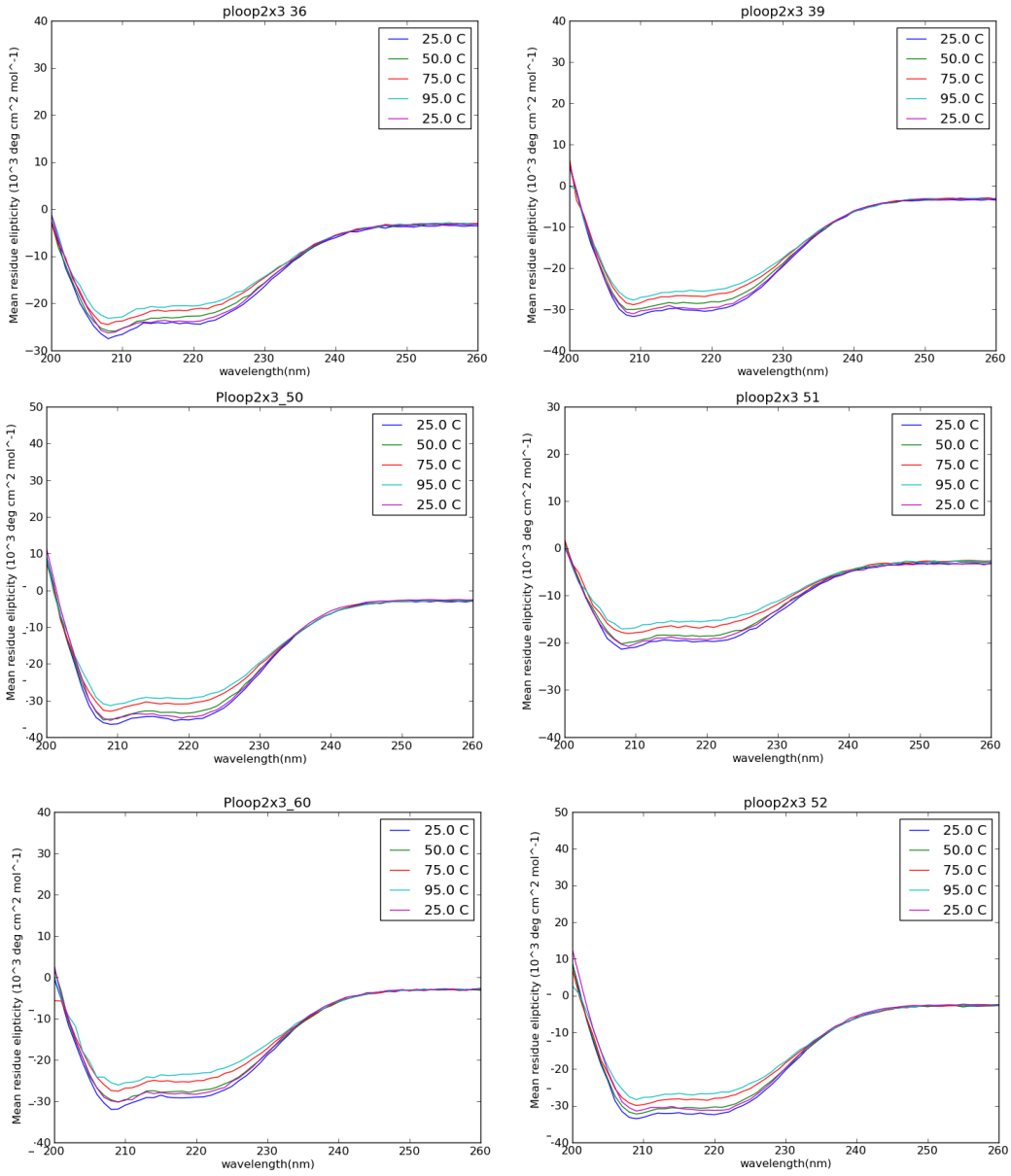


Figure 11 Circular Dichroism spectra of the tested designs. An initial scan was done for each protein at 25ºC (blue line) and then additional scans where performed at 50ºC (green line), 75ºC (red line) and 95ºC (cyan line). After the scan at 95ºC the samples were cooled down back to 25ºC and an additional scan was taken at 25ºC (purple line)

Five designs produced circular dichroism spectra compatible with a a αβ-protein and those where send to the collaborators at the *Northeast Structural Genomics consortium* (NESG) for H–^{15}N HSQC NMR analysis and structure determination.

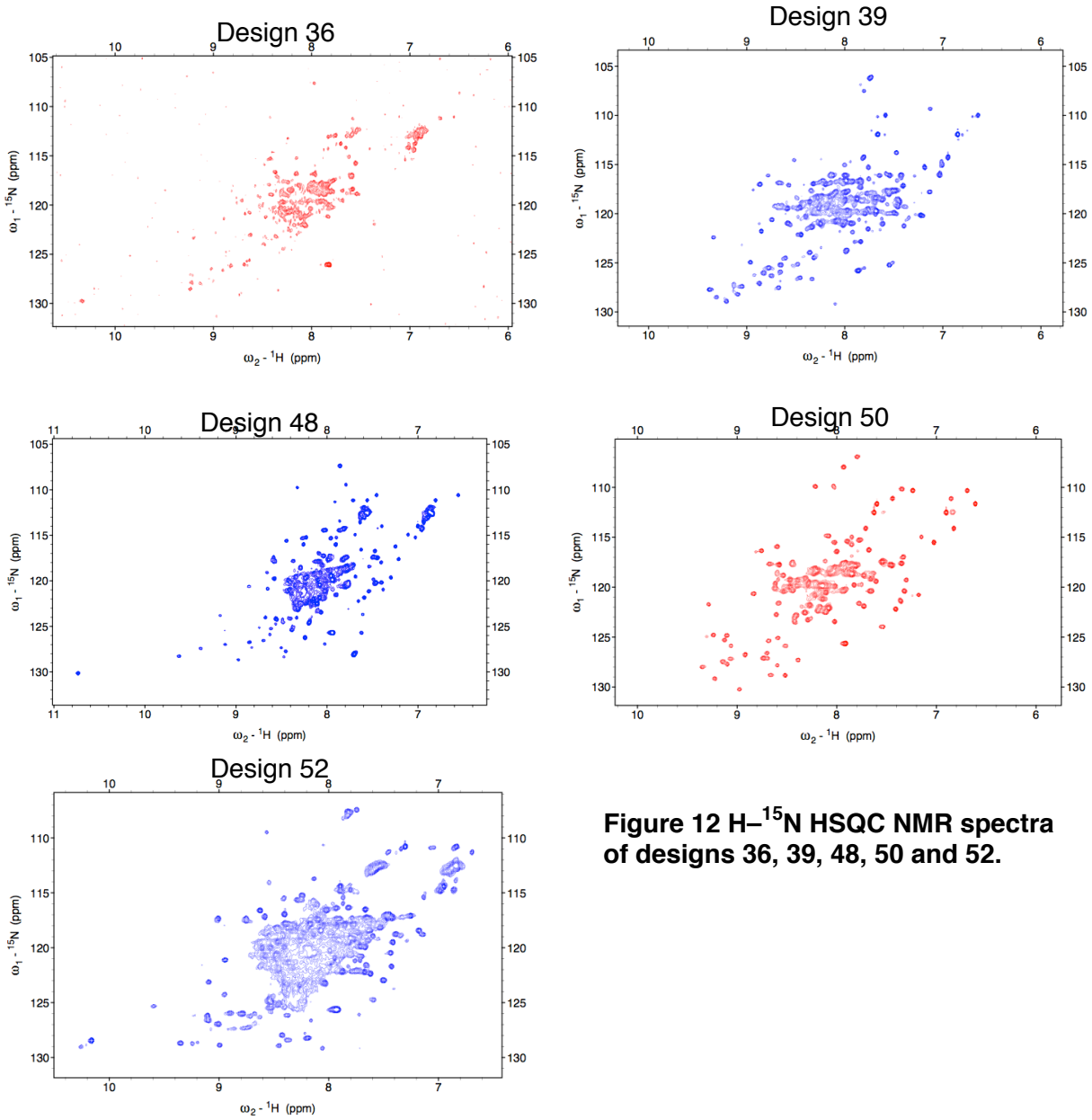
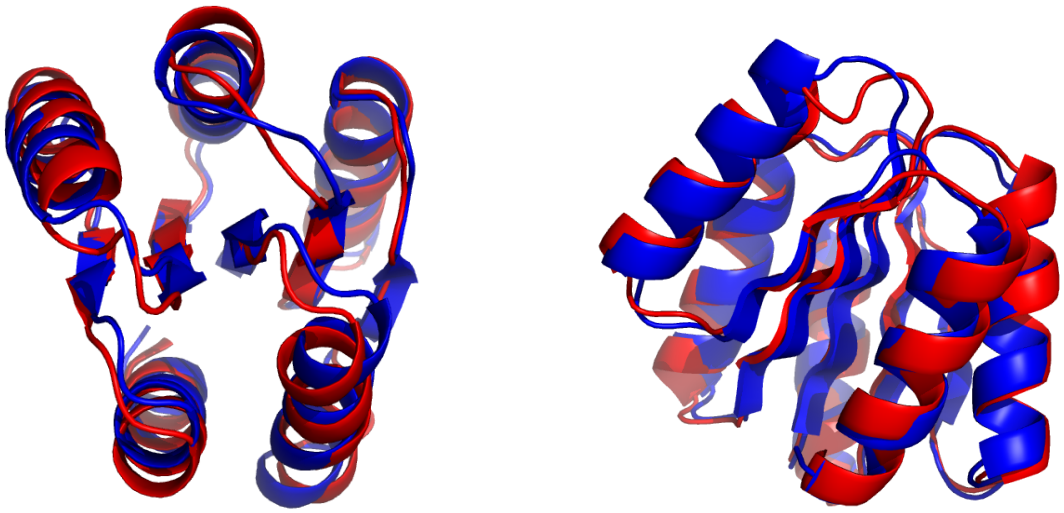


Figure 12 H–^{15}N HSQC NMR spectra of designs 36, 39, 48, 50 and 52.



NOE-based distance constraints				
Intraresidue			786	
Sequential			862	
medium range [1 < | i - j | < 5]			846	
long range [| i - j | ≥ 5]			1107	
Hydrogen bond constraints			68	
Dihedral-angle constraints			212	
Total			3881	
Structural Statistics				
Distance violations per model (Calculated using sum over r^-6)		0.1 - 0.2 Å	0.2 - 0.5 Å	0.5 Å
		6.25	0.7	0
Dihedral angle violations per model			1 - 10 °	> 10 °
			9.35	0
RMSD			*All backbone atoms*	*All heavy atoms*
			1.8 Å	2.4 Å
Molprobity analysis	Ramachandran plot statistics	*Most favoured regions*	*Allowed regions*	*Disallowed regions*
		98.2%	1.7%	0
	Clashscore	*Raw score*	*Z-score*	
		20.08	-1.92	

Figure 13. *top:* Cartoon representation of the design structure (blue) and NMR structure (red)**.** *bottom:* NMR structure statistics calculated among 20 refined structures.

Design 50 had the highest number of well defined peaks on the H–^{15}N HSQC (Figure 12) and so it was the most promising for NMR structure determination. The NMR structure of design 50 (Figure 13) shown that it was adopting the desired fold accurately (1.12Å Cα-RMSD to the designed model and a 2.40Å full-atom RMSD) with only 8 out of the 62 residues that form the core of the protein in rotamer different to the one of the design model.

Discussion

By following the rules described by Koga and Baker[27] it is possible to design novel α+β proteins with a high degree of accuracy. The rules are extremely useful for the design of loop regions but it is still necessary to do combinatorial sampling for strands and helices in order to get geometrically consistent backbones. It is necessary, and highly time consuming to the designer, to design against alternative conformations that might arise from rearrangements of the secondary structure elements. This can be achieved by carefully selecting the length of the secondary structure elements and through tight packing of the designed protein core in the desired conformation. The next and final chapter describes a new computational protocol that tries to solve this problem by performing iterative multistate negative design.

Methods

Protein Expression and Purification

All designed sequences were ordered from *Genescript* cloned into pET21b vector with a 6xHis tag at the C terminus. The proteins were expressed in *Escherichia coli* BL21 Star(DE3) cells. The cells were grown in MJ9 minimal media with ^{15}N ammonium sulphate as the sole nitrogen source and ^{12}C glucose as the sole carbon source. The cultures were grown overnight at 18ºC. Bacterial pellets were collected by centrifugation and cell lysis was done using BugBuster buffer (EMD Millipore). The soluble fraction of the lysate was collected by centrifugation and purified using nickel affinity columns. The columns where prepared by eluting 40 mL of binding buffer (PBS) before adding the samples.The columns where washed with 40 mL of PBS 30mM Imidazole, pH7.4 and the proteins where eluted with 12 mL of elution buffer(PBS). The purified proteins were then dialyzed overnight into PBS buffer.

Circular dichroism.

Circular dichroism data were collected on an Aviv 62A DS spectrometer. Far-ultraviolet circular dichroism spectra of designed proteins were measured from 260 to 200 nm for 14–25 mM protein samples in PBS buffer (pH 7.4) at various temperatures of 25, 50, 75 and 95 uC in a 1-mm-path-length cuvette. The protein concentrations were determined from the absorbance at 280 nm (ref. 50) using an ultraviolet spectrophotometer (NanoDrop, Thermo Scientific).

Size-exclusion chromatography combined with multi-angle light scattering. SEC-MALS experiments were performed using a miniDAWN TREOS static light-scattering detector (Wyatt Technology) combined with a HPLC system (LC 1200 Series, Agilent Technologies). One hundred microlitres of 400–700mM protein samples in PBS buffer (pH 7.4) was injected into a Superdex 75 10/300 GL column (GE Healthcare) equilibrated with PBS buffer at a flow rate of 0.5 ml min^{21}. The protein concentrations were calculated from the absorbance at 280 nm detected by the HPLC system. Static light-scattering data were collected at three different angles, 41.4u, 90.0u and 138.6u, at 658nm. These data were analyzed using ASTRA software (version 5.3.4, Wyatt Technology) with a change in the refractive index with concentration (dn/dc value) of 0.185 ml g^{21}.

Chapter Three: Iterative Multistate Negative Design of protein folds

Introduction

There has been considerable progress in the field of *de novo* protein design during the last two decades. Dahiyiat and Mayo designed a completely new sequence for a structure based on the zinc finger domain[11]. Harbury and Kim designed novel right-handed coil-coils from parametrically generated α-helices. [7]. Kuhlman and Baker successfully designed Top7, a *de novo* protein that has a new sequence and adopts a novel fold that hasn't been observed in nature[12]. More recently Koga and Baker have identified a set of rules for the design of *de novo* α+β proteins[27].

All of this designs have been made using the single state positive design approach. This design strategy consists on finding a sequence that minimizes the energy of the target structure. What is relevant for a protein fold is not the absolute energy of the target structure but the energy of the target structure relative to all the other accessible conformations[57]. Methods have been developed for multi state design that try to simultaneously stabilize the target state and destabilize a set of alternative states[58, 59]. These methods require a discrete set of states that need to be known in advance. For *ab initio* protein fold design such set of states is not know *a priori* and exhaustively enumerating all this states is unfeasible.

We developed a method that performs iterative mutistate negative design of a protein sequence by incorporating alternative conformations sampled using Rosetta's *ab initio* structure prediction protocol. The method works by identifying contacts between residues that are present in the alternative structures but not in the target fold. The residues that form these contacts are then scanned against a list of candidate mutations in an effort to maximize the energy gap between the designed state and the alternative state. The new optimized sequence is then evaluated by running it through the *ab initio* structure prediction protocol and the results are used in the next iteration. The new protocol can improve the energy landscape of the designs by improving the sampling in the region close to the design structure and eliminating or increasing the energy of alternative conformations.

Results

In the case of protein fold design a complete enumeration of all the possible conformations that a sequence might adopt is not feasible. The alternative strategy used in this work was to identify low energy competitor states and then optimize the protein sequence to simultaneously stabilize the target state and disfavor alternative states. This new protocol has to be run iteratively because changes in the protein sequence can alter the low energy competitor states

The new multistate optimization protocol is an iterative protocol that consists of the following steps:

1. Initial single state protein design.
2. Identify low energy competing states using *ab initio* protein structure prediction.
3. Analyze the contacts present in the low energy alternative states.
4. Perform multi-state negative design calculation over the residues identified in the previous step.
5. Return to step 2.

The crucial steps of the optimization protocol are step 3, identifying the problematic residues, and step 4, mutate those residues to destabilize the alternative states. A more thorough description of the algorithm can be found in the methods section.

Residue Contact Analysis

Contact matrices for the target fold and the alternative states are calculated from an energy landscape generated using the Rosetta *ab initio* structure prediction protocol(Figure 14A). From each decoy contact matrix the target contact matrix is subtracted the results are summed to generate the contact difference matrix shown in Figure 14.B. The areas in that are colored in blue in the contact difference matrix represent contacts that are consistently present in the decoy

structures but not in the target structure. The dark red areas represent contacts that are under sampled.

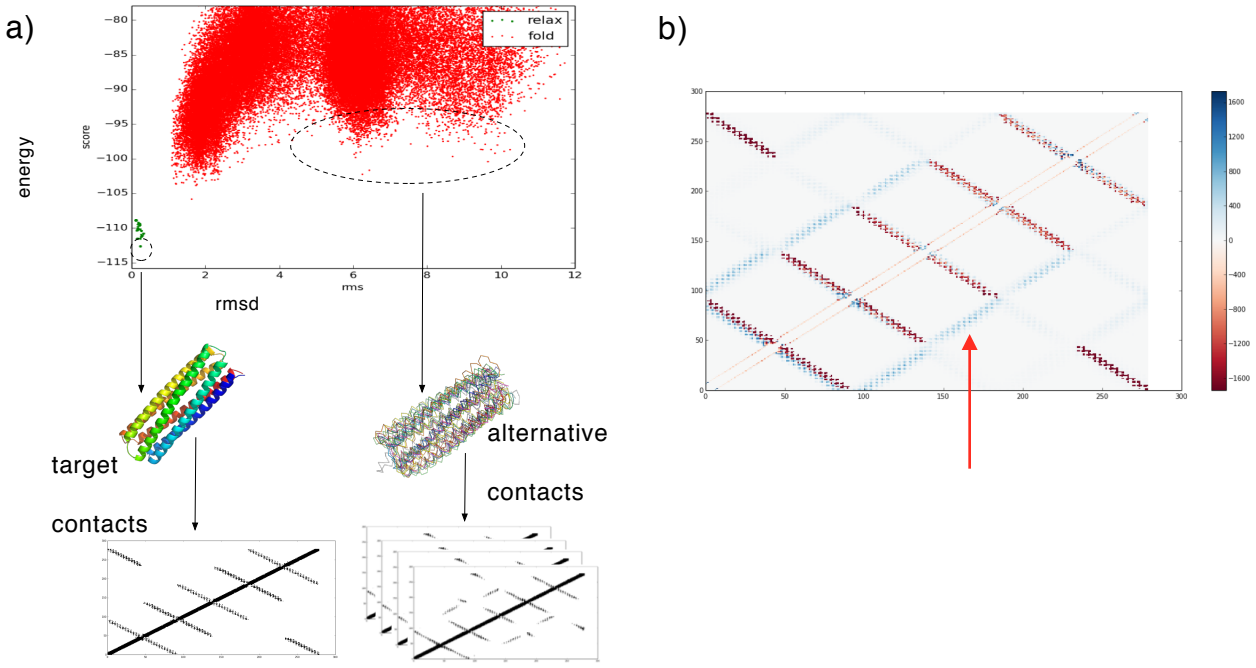


Figure 14. Illustration of contact analysis.

a) *top:* Example of energy landscape for a input sequence produced with the Rosetta protein structure prediction protocol. The x-axis is the root mean square deviation to the target structure and the y-axis is the energy in Rosetta energy units. *middle:* cartoon representation of the target and low energy alternative states. *bottom:* contact maps produced for the target and alternative states.

b) Contact difference map produced by subtracting the target contacts from the alternative contacts and adding up all the difference matrices. The areas in dark blue represent unwanted contacts between residues present in the alternative low energy states(e.g. area pointed by red arrow).

After the difference matrix is generated the pairs of residues that make unwanted contacts are sorted by their abundance and organized into a graph (Figure 15). Each group in the graph represents a cluster of unwanted interactions. The edges in the graph shown in solid lines represent contacts between residues that are present in more than 90% of the decoys.

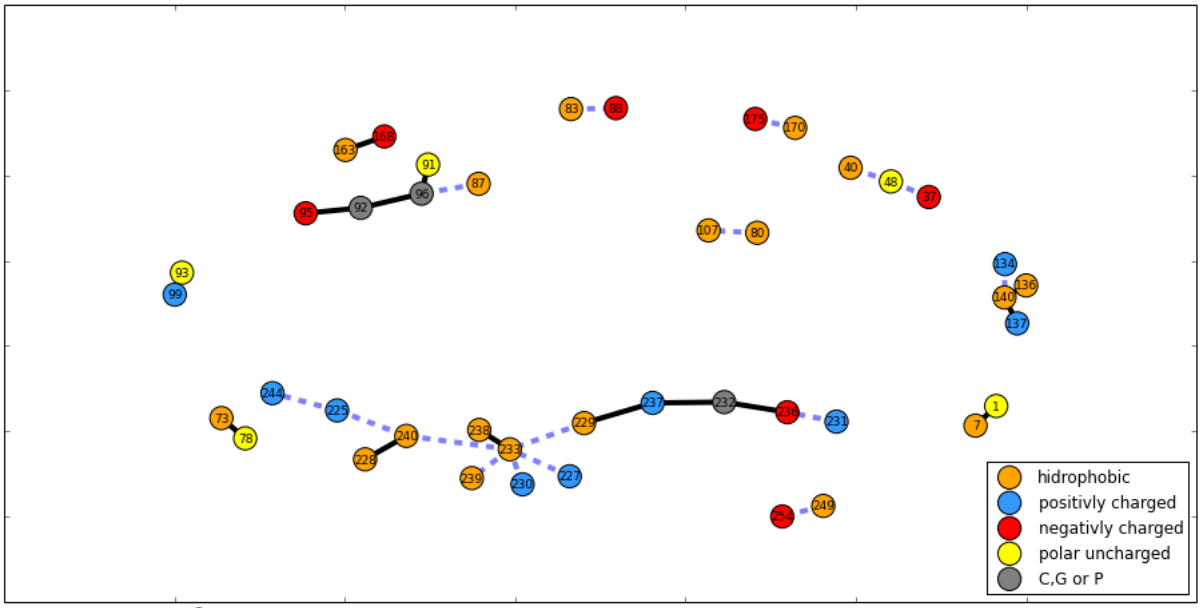


Figure 15. Contact analysis of the unwanted residue interactions present in the alternative low energy conformations. The nodes represent residue positions the analyzed protein, the edges represent the interaction between two residues. In solid lines are represented the contacts that appear in more than 85% of the low energy alternative states. In this particular example the are four interaction groups with more than two nodes. The residues in groups that have more than two elements and that have at least two edges connected to them are the ones that are selected to be optimized by the multistate negative design algorithm.

The nodes in each group are sorted by their number of edges and the nodes that have only one edge are removed. The remaining nodes are combined into a list

by iterating over the groups and picking the node with the highest degree from each group. This list contains the candidate residues that need to be mutated to avoid unwanted interactions. The list is sorted by degree because mutating a residue with high degree can disrupt many interactions simultaneously reducing the number of mutations necessary to avoid the alternative state-stabilizing interactions.

Multi-state negative design

The list of interactions compiled in the previous step is used to guide the multi-state negative design calculation. Each of the selected positions is scanned by simultaneously changing the side chain in the target and in the low energy decoys. At each position only mutations which are compatible with the backbone and and degree of burial of the residue are considered. The effect of each mutation is scored based on the energy gap between the target structure and the lowest energy alternative states (Figure 16). The score is a weighted sum of the difference in energy gap before and after applying the mutation. The mutation is accepted if the score is positive (the gap between the target and the decoys increased) and the energy of the target is lower or close to the energy of the initial structure. The optimization protocol uses a greedy approach optimizing first the residues with higher node degree.

Since each mutation introduced during the optimization process changes the energy landscape the decoys need to be recalculated by running the *ab initio*

structure prediction protocol over the new optimized sequence. The mutations introduced in each optimization step are limited to be more than two and less than ten to ensure changes in the energy landscape and at the same time avoid drifting too much from the input sequence.

$$\Delta E(aa, pos) = [\text{mut\_test\_score(aa,pos)} - \text{alt\_state\_init\_score}] - [\text{target\_test\_score(aa,pos)} - \text{target\_init\_score}]$$

$$score(aa, pos) = \frac{\sum_i \Delta Ei_{(aa,\,pos)} * e^{\frac{-\Delta Ei(aa,pos)}{kt}}}{\sum e^{\frac{-\Delta Ei(aa,pos)}{kt}}}$$

Figure 16. Score function used during the multistate negative design. ΔE(aa, pos) is the energy difference between the alternative state and the target state when mutation *aa* is introduced at position *pos.* The score(aa,pos) is a weighted sum score the ΔE(aa, pos) calculated over all alternative states *i.*

Test cases

Ferredoxin fold

The ferredoxin fold is an α+β fold composed of 4 strands and two helices. A smaller version of an already existing ferredoxin was generated by changing the loop regions for new shorter ones from fragments of proteins in the PDB. The standard protocol for sequence design was then run over the new backbone and the newly generated sequence was then evaluated by running it through the *ab initio* structure prediction protocol. The resulting energy landscape for the design

sequence showed an alternative minima around 10 Å rms to the target design that was heavily sampled(Figure 17, top left).

Two iterations of the multistate negative design protocol eliminated the alternative minima and increased the sampling in the region close to the target. The new generated energy landscape has a narrower and better defined funnel when compared to the original.

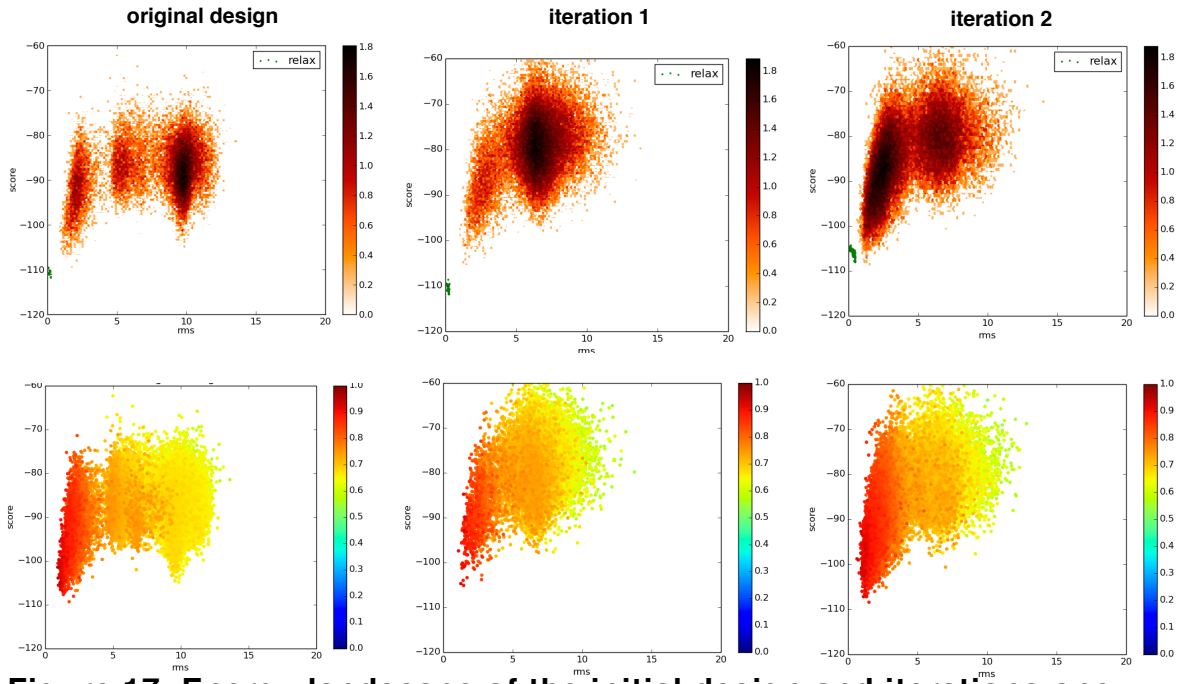


Figure 17. Energy landscape of the initial design and iterations one and two for the ferredoxin fold. The x-axis represents the RMSD to the target structure and the y-axis represent score in Rosetta energy units. *top:* Energy landscape colored by density, darker regions represent zones of the landscape with higher sampling. *bottom*: Energy landscape colored by fraction of contacts present in the target structure.

Two-Helix Bundle

In this example the original design is composed of two alpha helices connected by a 3 residue loop that pack against each other. This structure was generated in a puzzle by the Foldit players[60]. The energy landscape of the original design was primarily concentrated in two clusters around 5 and 7 Å rms(Figure 18). After the first iteration the sampling improves in the region below 5Å rms and by the fifth iteration most of the sampling is concentrated around 3Å rms with a few samples below 1Å rms. The sampling improved considerably by the fifth iteration generating a smooth energy funnel towards the designed structure.

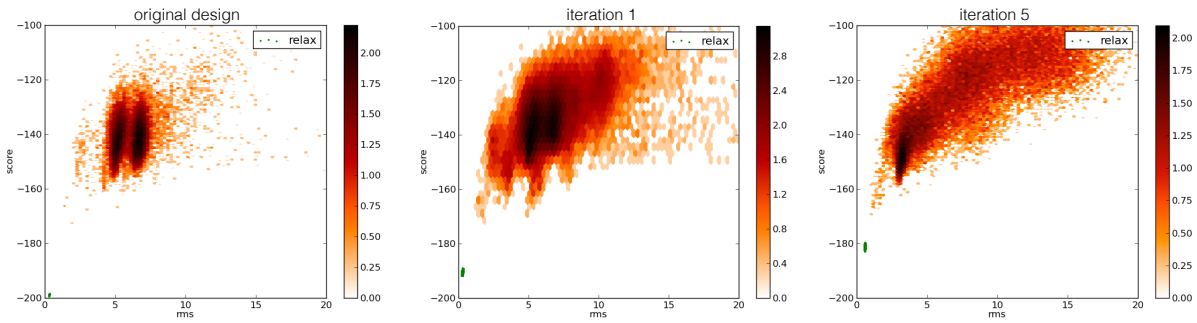


Figure 18. Energy lansdcape of the initial design and iterations one and five for the two-helix bundle. The x-axis represents the RMSD to the target structure and the y-axis represent score in Rosetta energy units. *top:* Energy landscape colored by density, darker regions represent zones of the landscape with higher sampling.

Three-Helix Bundle

This example was also generated by Foldit players and consists of a single chain three-helix bundle. In the original design the sampling was highly concentrated in the area between 9-10Å rms(Figure 19). After five iterations the energy of the sampling improves increasing significantly in the area below 2Å rms and by the tenth iteration most of the sampling is concentrated in this region.

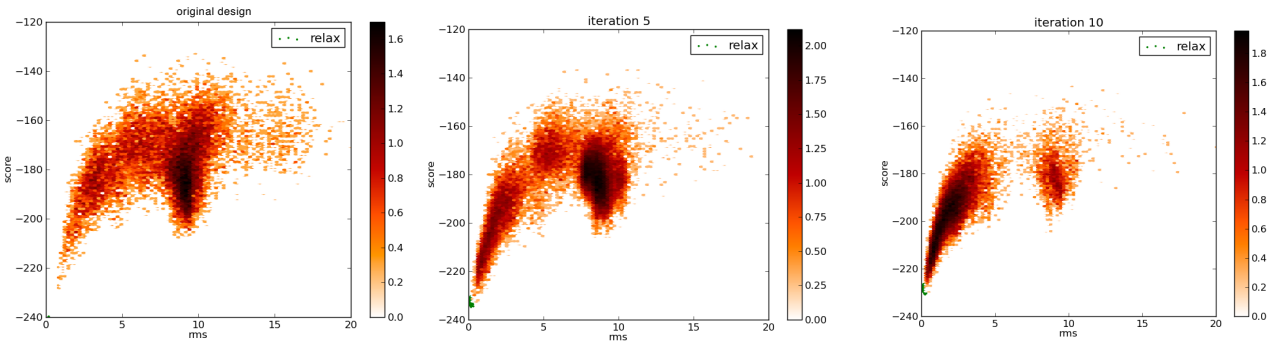


Figure 19. Energy lansdcape of the initial design and iterations five and ten for the three-helix bundle. The x-axis represents the RMSD to the target structure and the y-axis represent score in Rosetta energy units. *top:* Energy landscape colored by density, darker regions represent zones of the landscape with higher sampling.

P-loop fold

The P-loop fold is an α+β Rossmann-like fold composed of 5 helices and 5 strands in 54132 orientation. In the original design the sampling is concentrated in a cluster around 10Å. The structures in this cluster contain many of the contacts present in the designed structure suggesting that the design is partially folded. After the first iteration some of the contacts that support the cluster are destabilized and the sampling is increased in the area around 5Å rms. By the third iteration the sampling has improved even further with samples as low as 2Å rms from the designed structure.

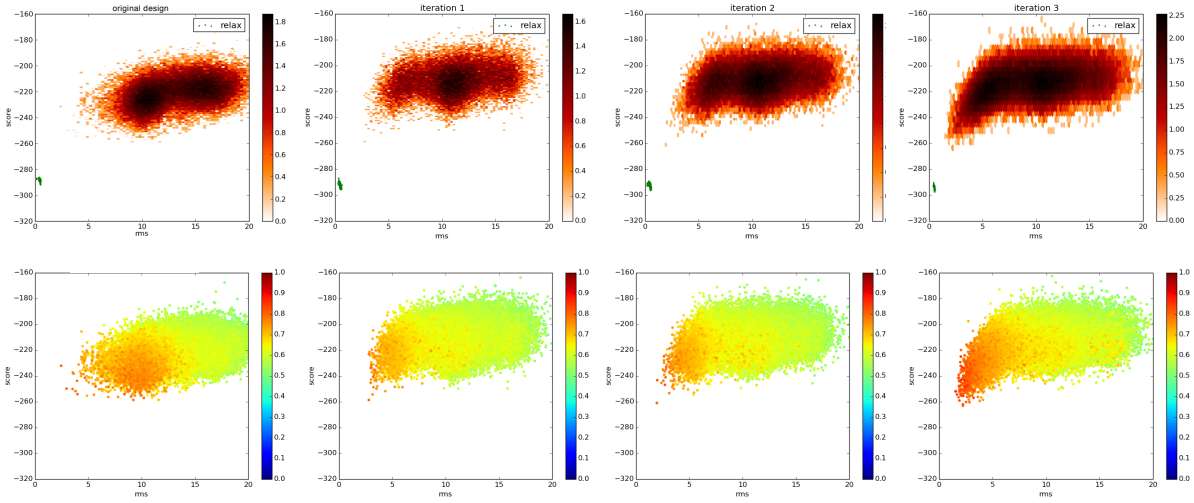


Figure 20. Energy lansdcape of the initial design and iterations one, two and three for the P-loop fold. The x-axis represents the RMSD to the target structure and the y-axis represent score in Rosetta energy units. *top:* Energy landscape colored by density, darker regions represent zones of the landscape with higher sampling. *bottom*: Energy landscape colored by fraction of contacts present in the target structure.

Discussion

Iterative multi state negative design can effectively improve the energy landscape of *de novo* design proteins by eliminating unwanted interactions that might form during the folding of the protein. This new method should increase the chances of a design to fold to the correct structure by destabilizing the alternative states. This hypothesis can be tested by using the protocol on a set of targets topologies designed using the standard RosettaDesign protocol and then comparing the experimental results of optimized proteins against unoptimized.

The protocol captures many aspects of the sequence maturation process that are usually done manually by the designer. It is able to identify and fix problems in designs that are difficult and time consuming to discover. One of the advantages of this new method that it is able correct small inconsistencies present in the backbone due to unrealistic geometry. Instead of running the expensive *ab initio* structure prediction calculation on a large sample of designed proteins expecting the designer, using the iterative protocol, can ran few designs that will be 'matured' during the iterations and in many cases generating a consistent energy landscape.

The strategy defined in this protocol, iterative identification of problematic residues and multi state negative design, can be extended to solve other problems different from *de novo* fold design. Iterative multi state negative design

can be used during the design of protein-protein surface interactions by changing the sampling method from *ab initio* structure prediction to global docking and doing the contact analysis on the residues on the predicted interface. The protocol could also be applicable to ligand docking design by using the same logic.

Methods

Single state sequence design

Single state sequence design was performed using the flexible backbone and Solvent Accessible Surface Area (SASA) layers strategy[27].
This method consists in iterations of fixed backbone sequence design with the RosettaDesign algorithm[12] followed full atom minimization using the FastRelax algorithm[61].

Iterative multistate design protocol

The new protocol was implemented using the Pyrosetta[62] programming interface for the Rosetta3 software suite[63].

The new multistate protocol consists of the following steps:

1. Initial single state protein design.
2. Sampling of the alternative states using *ab initio* protein structure prediction.

3. Extraction of the lowest thousand energy structures from the sampled decoys.
4. Residue contact analysis of the low energy alternative states.
5. Identification the small subset of residues that are mostly responsible for the unwanted contacts present on the alternative states.
6. Assemble lists of candidate mutations for the selected residues by analysis the surface exposure of those residues.
7. Reduce the candidate mutations to those that are compatible with the backbone of the target by fragment analysis.
8. Perform multi-state negative design calculation over the residues identified in the previous step.
9. Return to step 2.

The *ab initio* structure prediction protocol (step 2) used in this work was run on the Rosetta@home distributed platform as described[65].

During the multistate negative design(step 7) mutation are rejected if they destabilize the input structure by more than 10 Rosetta energy units.

Bibliography

1. Drexler, K.E., Molecular engineering: An approach to the development of general capabilities for molecular manipulation. Proceedings of the National Academy of Sciences, 1981. 78(9): p. 5275-5278.

2. Ponder, J.W. and F.M. Richards, Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of Molecular Biology, 1987. 193(4): p. 775-791.

3. Hellinga, H.W. and F.M. Richards, Optimal sequence selection in proteins of known structure by simulated evolution. Proceedings of the National Academy of Sciences, 1994. 91(13): p. 5803-5807.

4. Desjarlais, J.R. and T.M. Handel, De novo design of the hydrophobic cores of proteins. Protein Science, 1995. 4(10): p. 2006-2018.

5. Lazar, G.A., J.R. Desjarlais, and T.M. Handel, De novo design of the hydrophobic core of ubiquitin. Protein science : a publication of the Protein Society, 1997. 6(6): p. 1167-1178.

6. Bryson, J.W., et al., From coiled coils to small globular proteins: design of a native-like three-helix bundle., in Protein Sci. 1998, Cold Spring Harbor Laboratory Press. p. 1404-1414.

7. Harbury, P.B., et al., High-Resolution Protein Design with Backbone Freedom. Science, 1998. 282(5393): p. 1462-1467.

8. Hill, R.B., et al., De NovoDesign of Helical Bundles as Models for Understanding Protein Folding and Function. Accounts of Chemical Research, 2000. 33(11): p. 745-754.

9. Lasters, I., M. De Maeyer, and J. Desmet, Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. Protein engineering, 1995. 8(8): p. 815-822.

10. Gordon, D.B. and S.L. Mayo, Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. Journal of Computational Chemistry, 1998.

11. Dahiyat, B.I. and S.L. Mayo, De Novo Protein Design: Fully Automated Sequence Selection. Science, 1997. 278(5335): p. 82-87.

12. Kuhlman, B., et al., Design of a Novel Globular Protein Fold with Atomic-Level Accuracy, in Science. 2003, American Association for the Advancement of Science. p. 1364-1368.

13. Michalopoulos, I., et al., TOPS: an enhanced database of protein structural topology. Nucleic acids research, 2004. 32(suppl 1): p. D251-D254.

14. Bowers, P.M., C.E. Strauss, and D. Baker, De novo protein structure determination using sparse NMR data. Journal of biomolecular NMR, 2000. 18(4): p. 311-318.

15. Kuhlman, B. and D. Baker, Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences, 2000. 97(19): p. 10383-10388.

16. Bolon, D.N. and S.L. Mayo, Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences, 2001. 98(25): p. 14274-14279.

17. Kaplan, J. and W.F. DeGrado, De novo design of catalytic proteins. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(32): p. 11566-11570.

18. Zanghellini, A., et al., New algorithms and an in silico benchmark for computational enzyme design. Protein Science, 2006. 15(12): p. 2785-2794.

19. Jiang, L., et al., De novo computational design of retro-aldol enzymes. science, 2008. 319(5868): p. 1387-1391.

20. Röthlisberger, D., et al., Kemp elimination catalysts by computational enzyme design. Nature, 2008. 453(7192): p. 190-195.

21. Siegel, J.B., et al., Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science, 2010. 329(5989): p. 309-313.

22. Kortemme, T. and D. Baker, A simple physical model for binding energy hot spots in protein–protein complexes. Proceedings of the National Academy of Sciences, 2002. 99(22): p. 14116-14121.

23. Shifman, J.M. and S.L. Mayo, Modulating calmodulin binding specificity through computational protein design. Journal of molecular biology, 2002. 323(3): p. 417-423.

24. Huang, P.S., J.J. Love, and S.L. Mayo, A de novo designed protein–protein interface. Protein Science, 2007. 16(12): p. 2770-2774.

25. King, N.P., et al., Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science, 2012. 336(6085): p. 1171-1174.

26. King, N.P., et al., Accurate design of co-assembling multi-component protein nanomaterials. Nature, 2014. 510(7503): p. 103-108.

27. Koga, N., et al., Principles for designing ideal protein structures., in Nature. 2012. p. 222-227.

28. Grigoryan, G. and W.F. DeGrado, Probing Designability via a Generalized Model of Helical Bundle Geometry. Journal of Molecular Biology, 2011. 405(4): p. 1079-1100.

29. Simons, K.T., et al., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions., in Journal of Molecular Biology. 1997. p. 209-225.

30. Abagyan, R. and M. Totrov, Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. Journal of molecular biology, 1994. 235(3): p. 983-1002.

31. Kolinski, A. and J. Skolnick, Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins: Structure, Function, and Bioinformatics, 1994. 18(4): p. 338-352.

32. Irbäck, A., et al., Monte Carlo procedure for protein design. Physical Review E, 1998. 58(5): p. R5249.

33. Lasters, I. and J. Desmet, The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. Protein engineering, 1993. 6(7): p. 717-722.

34. Maeyer, M.D., J. Desmet, and I. Lasters, All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Folding and Design, 1997. 2(1): p. 53-66.

35. Malakauskas, S.M. and S.L. Mayo, Design, structure and stability of a hyperthermophilic protein variant. Nature structural biology, 1998. 5(6): p. 470-475.

36. Pierce, N.A., et al., Conformational splitting: A more powerful criterion for dead-end elimination. Journal of computational chemistry, 2000. 21(11): p. 999-1009.

37. Looger, L.L. and H.W. Hellinga, Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. Journal of molecular biology, 2001. 307(1): p. 429-445.

38. Ben-Naim, A., Statistical potentials extracted from protein structures: Are these meaningful potentials? The Journal of chemical physics, 1997. 107(9): p. 3698-3706.

39. Mohanty, D., et al., Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper. Proteins: Structure, Function, and Bioinformatics, 1999. 35(4): p. 447-452.

40. Lazaridis, T. and M. Karplus, Effective energy functions for protein structure prediction. Current opinion in structural biology, 2000. 10(2): p. 139-145.

41. Boas, F.E. and P.B. Harbury, Potential energy functions for protein design. Current opinion in structural biology, 2007. 17(2): p. 199-204.

42. Brooks, B.R., et al., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of computational chemistry, 1983. 4(2): p. 187-217.

43. Pearlman, D.A., et al., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Computer Physics Communications, 1995. 91(1): p. 1-41.

44. Sippl, M.J., Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins. Journal of Molecular Biology, 1990. 213(4): p. 859-883.

45. Kortemme, T., A.V. Morozov, and D. Baker, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. Journal of molecular biology, 2003. 326(4): p. 1239-1259.

46. Dahiyat, B.I. and S.L. Mayo, Protein design automation. Protein Science, 1996. 5(5): p. 895-903.

47. Dunbrack Jr, R.L. and M. Karplus, Backbone-dependent Rotamer Library for Proteins Application to Side-chain Prediction. Journal of Molecular Biology, 1993. 230(2): p. 543-574.

48. Metropolis, N., et al., Equation of state calculations by fast computing machines. The journal of chemical physics, 1953. 21(6): p. 1087-1092.

49. Petrey, D., et al., Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins: Structure, Function, and Bioinformatics, 2003. 53(S6): p. 430-435.

50. Schwede, T., et al., SWISS-MODEL: an automated protein homology-modeling server. Nucleic acids research, 2003. 31(13): p. 3381-3385.

51. Eswar, N., et al., Comparative protein structure modeling using Modeller. Current protocols in bioinformatics, 2006: p. 5.6. 1-5.6. 30.

52. Chothia, C. and A.M. Lesk, The relation between the divergence of sequence and structure in proteins. The EMBO journal, 1986. 5(4): p. 823.

53. Principles that govern the folding of protein chains, in Science. 1973.

54. Simons, K.T., et al., Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Structure, Function, and Bioinformatics, 1999. 37(S3): p. 171-176.

55. Zhang, Y., A. Kolinski, and J. Skolnick, TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophysical journal, 2003. 85(2): p. 1145-1164.

56. Sheffler, W. and D. Baker, RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Science, 2009. 18(1): p. 229-239.

57. Berezovsky, I.N., K.B. Zeldovich, and E.I. Shakhnovich, Positive and negative design in stability and thermal adaptation of natural proteins., in PLoS Comput. Biol. 2007, Public Library of Science. p. e52.

58. Leaver-Fay, A., et al., A generic program for multistate protein design., in PLoS ONE. 2011, Public Library of Science. p. e20937.

59. Multistate approaches in computational protein design, in Protein Science. 2012, Wiley Subscription Services, Inc., A Wiley Company. p. 1241-1252.

60. Cooper, S., et al., Predicting protein structures with a multiplayer online game. Nature, 2010. 466(7307): p. 756-760.

61. Tyka, M.D., et al., Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. Journal of Molecular Biology, 2011. 405(2): p. 607-618.

62. Chaudhury, S., S. Lyskov, and J.J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta., in Bioinformatics. 2010, Oxford University Press. p. 689-691.

63. Leaver-Fay, A., et al., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol, 2011. 487: p. 545-574.

64. Sancho, J., Flavodoxins: sequence, folding, binding, function and beyond. Cellular and Molecular Life Sciences CMLS, 2006. 63(7-8): p. 855-864.

65. Das, R., et al., Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home., in Proteins. 2007. p. 118-128.

VITA

Javier Castellanos was born in Santiago, Chile and currently lives in Seattle, Washington. He attended the Andree English School in Santiago. He received a bachelors degree in Biochemistry from University of Chile, Santiago. In 2014 he received a Doctor of Philosophy degree from the University of Washington.