COMPUTATIONAL DESIGN OF NOVEL PROTEINS FOR AFFINITY REAGENTS

Ramesh Kumar Jha

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics (Program in Molecular and Cellular Biophysics).

Chapel Hill 2010

Approved by:
Advisor: Brian Kuhlman
Reader: Sharon Campbell
Reader: Klaus Hahn
Reader: Hengming Ke
Reader: Garegin Papoian
ABSTRACT

Ramesh Kumar Jha: Computational design of novel proteins for affinity reagents
(Under the direction of Brian Kuhlman)

Computational protein design is an emerging technology that can pave a way to redesign and create novel proteins which have new functions as affinity reagents and protein therapeutics. Two approaches have been described to create an affinity reagent for activated form of p21-activated kinase 1 (PAK1). A good affinity reagent can be developed into a biosensor for visualizing the activated form of PAK1 inside cells.

In the first approach, truncated auto-inhibitory domain (trunc-AID) from PAK1, which is unfolded but binds the activated conformation of PAK1 with an affinity of 4 µM, was redesigned for improvement in structure and function. Using a computational approach implemented within Rosetta molecular modeling program, a 20-residue helix at the N-terminus was created in the N-terminus of trunc-AID that had the PAK1-binding region conserved. The design, AlmostHelix showed gain in structure and solubility but no improvement in binding affinity.

Another design, PAcKer, where a 16-residue helix was created at the C-terminus of the trunc-AID and was inserted in a fluorescent protein (CFP) bound the active conformation of PAK1 with an affinity of ~400 nM. CFP-PAcKer did not show any binding to the ‘closed’ form of PAK1 making it suitable for a biosensor that can detect only the activated form of PAK1 inside cells.
A de novo interface design approach was used to create a binding interface on a segment from hyperplastic discs protein (‘scaffold’) to bind the active conformation of PAK1 (‘target’). Using a newly developed protocol DDMI, that is implemented within the Rosetta molecular modeling program and uses rigid-body docking, sequence design, and gradient-based minimization of backbone and side chain torsion angles to design low energy interfaces between the ‘scaffold’ and ‘target’, a binder called Spider Roll was designed. Spider Roll bound to PAK1 with a modest affinity of 100 µM. Mutagenesis studies confirmed that the binding interface was consistent with the design model. NMR studies were also consistent with the binding model. Additionally, Spider Roll did not bind the ‘closed’ PAK1. De novo designed binders that can distinguish between two states of proteins have potential use as biosensors.

As a future direction for de novo protein interface design, an attempt was made to create a metal-mediated interaction between a ‘scaffold’ protein and Ubiquitin as a ‘target’. Metal-mediated interactions have advantage over other non-covalent interactions due to strong coordination bond. Experimental testing of multiple designs showed a design (called Spelter) bound with low nanomolar affinity to the metal ion, but any metal-mediated interaction between Spelter and Ubiquitin was not established.
Dedicated
To my lovely wife and cute little son for their inspiration
&
To all my colleagues, friends and family members for their support
&
To all my teachers for everything they taught me
ACKNOWLEDGEMENTS

I would like to thank Brian for allowing me to pursue my doctoral dissertation under his guidance. During my research work he has been a wonderful mentor with enormous amount of patience. He taught me how ‘successes’ and ‘failures’ are both important part of research and how we can learn from both.

I would also like to thank Andrew Leaver-Fay and a former postdoc Glenn Butterfoss for laying the foundation for protein interface design and to Steven Lewis and Bryan Der for their important contribution in the metal-mediated interface design project. I also convey my deep sense of gratitude for all present and past members of Kuhlman lab and Rosetta community for creating such a nice environment for research and helping me overcome the day to day hurdles. I would specially like to thank Carrie, Gurkan, Eun-Jung, Doug, Jenny, Ron, Grant, Ben, Oana and Matt for their stimulating discussions during work and in the lab meetings.

I would like to thank my collaborators, Klaus and Yi for giving me an opportunity to design biosensor against PAK1 and for doing in vivo studies. Throughout my work I had very fruitful discussions without which the work would not have been possible. I would also like to thank Thomas and Yibing (SUNY Buffalo) for assigning the NMR chemical shifts of Spider Roll. I would like to extend my gratitude towards Nikolay and Shuangye for carrying DMD simulations on Spider Roll and towards Hengming and Huanchen for helping me set up crystallization trays for Spider Roll. I would also like to thank Maren and Don Hilvert (ETH, Zurich) for carrying directed evolution on Spider Roll.
I am also grateful to Ash (Tripathy) from Macinfac and Greg (Young) from NMR facility for their kind support and fruitful discussions during analyses of my experimental data.

I would like to thank all my committee members for their support and fruitful discussions during my committee meetings and for their availability for any kind of discussions related to my research.

I convey my special thanks to my grand-parents, parents, brother and sister for their encouragement and to my wife Rashmi for her inspiration during my PhD program and to my son Suchir for his entertainment during the time of dissertation writing.

Last but not least I thank the Department of Biochemistry and Biophysics and Program of Molecular and Cellular Biophysics for giving me opportunity to work in an incredible intellectual environment.
# TABLE OF CONTENTS

LIST OF TABLES.............................................................................................................................................xi

LIST OF FIGURES...........................................................................................................................................xii

LIST OF ABBREVIATIONS.............................................................................................................................xv

Chapter

1. INTRODUCTION...........................................................................................................................................1

   Recent development in computational protein design.................................................................1
   Computational protein design for functional improvements..................................................2
   Previous works in de novo design of protein-protein interface........................................4
   Rosetta molecular modeling suite.........................................................................................5
   Affinity reagents for biosensors.............................................................................................8
   Model system, goal and design approach.............................................................................9
   Incorporating metal ions in the protein interface as a future direction .......................14
   References..............................................................................................................................16

2. COMPUTATIONAL REDesign OF PAK1 AUTO-INHIBITORY DOMAIN FOR A BIOSENSOR........................................................................................21

   Abstract...............................................................................................................................................21
   Introduction.......................................................................................................................................23
3. COMPUTATIONAL DESIGN OF A PAK1 BINDING PROTEIN……..……..52
   Abstract………………………………………………………………….52
   Introduction…………………………………………………………….54
   Results…………………………………………………………………56
   Discussions………………………………………………………………73
   Materials and methods………………………………………………….75
   Supplementary information…………………………………………83
   References………………………………………………………………94

4. DESIGN OF A METAL-MEDIATED PROTEIN-PROTEIN INTERACTION…………………………………………………………………..…….98
   Abstract…………………………………………………………………98
   Introduction………………………………………………………………99
   Results……………………………………………………………………101
   Discussions……………………………………………………………..109
   Materials and methods…………………………………………………111
   Supplementary information…………………………………………116
   References………………………………………………………………119
5. CONCLUSIONS AND FUTURE DIRECTIONS………………………………..….121

Redesign of PAK1 auto-inhibitory domain for a biosensor…………………………….121

De novo design of a PAK1 binder…………………………………………………………123

References……………………………………………………………………………………127
LIST OF TABLES

Table

2.1 Biophysical properties of designs and comparison with the truncated AID .............................................................. 31

3.1 Benchmark scores from native PDBs and the cutoffs used as Filter II during simulations ............................................. 61

3.2 Computational scores of the selected design models and comparison with a native AID-kinase interaction in PAK1 full length ........................................................................................................ 63

3.3 Biophysical characterization of the selected designs .......................................................... 64

3.4 Buried solvent accessible surface area (SASA) of the interface residues in Spider Roll – kinase complex in two alternate docked positions ........................................................................................ 71

S3.1 High resolution PDBs used for evaluation of DDMI protocol ........................................ 88

S3.2 Backbone assignment of Spider Roll I58A at 288 K ............................................................................. 89

S3.3 Length and composition comparison of the secondary structure elements between X-ray structure and CS-Rosetta structure ........................................................................................................ 91

4.1 Tetrahedral geometry parameters for Zn-coordination .................................................. 103

4.2 Computational scores of the design models based on which selections were made ........................................................................ 110

S4.1 List of scaffolds used for Zn-mediated de novo interface design ........................................ 117

S4.2 Amino acid sequence of the designs selected ........................................................................ 118
LIST OF FIGURES

Figure

1.1 Proposed function of an affinity reagent .................................................. 9
1.2 Activation mechanism of p21-activated kinase 1 (PAK1) ..................... 10
1.3 Crystal structure of auto-inhibited PAK1 ............................................. 11
1.4 Structural alignment of PAK1 kinase domain and kinase domain from auto-inhibited PAK1 .............................................................. 12
1.5 The two approaches for designing affinity reagents .......................... 13
2.1 Structural and primary sequence comparison of the models evolved from the auto-inhibitory domain (AID, cyan) .............................. 27
2.2 Structural characterization of computational models and WT domain using CD .......................................................... 30
2.3 Isothermal titration calorimetry assays to determine the binding affinities of CFP-PAcKer against active and auto-inhibited conformations of PAK1 ............................................. 32
2.4 Biosensor derived from CFP-PAcKer selectively ‘lights up’ in the presence of PAK1 kinase domain .................................................. 34
2.5 CFP-PAcKer derived biosensor gives a poor response with PAK5 kinase domain .......................................................... 35
S2.1 AlmostHelix gains structure in the presence of TMAO ....................... 43
S2.2 Isothermal titration calorimetry assay to determine the binding affinity of CFP-PAcKer against ‘open’ form of PAK1 ......................... 44
S2.3 A destructive mutation on either side of proposed CFP-PAcker and kinase domain interaction resulted in a complete loss of signal in ITC .................................................. 45
S2.4 Mutant biosensor (CFP-PAcKer^{V127E} based) against PAK1 kinase domain .......................................................... 46
S2.5 ITC assay to detect interaction between CFP-PAcKer and GST-PAK5 kinase domain .......................................................... 46
S2.6 GST-PAK5 kinase shows poor interaction with the CFP-PAcKer derived biosensor but shows a different saturation.........................47

S2.7 ITC assay to show CFP-IS and PAK1 kinase domain interaction.................................................................48

3.1 DDMI protocol for protein interface design..........................................................58

3.2 Conformational sampling in ‘dock’ stage and energy convergence during the ‘design/minimize’ stage............................59

3.3 Multiple sequence alignment of the design models selected for experimental validation.........................................................62

3.4 Spider Roll – PAK1 model showing the interface residues..................65

3.5 Mutagenesis studies of Spider Roll - PAK1 interface..........................66

3.6 Normalized Spider Roll I58A HSQC peak volumes as a function of PAK1 concentration.........................................................70

3.7 Presence of two competing states in Spider Roll – PAK1 model.............72

3.8 Spider Roll distinguishes between ‘closed’ and ‘open’ full length PAK1 (PAK1-fl).........................................................73

S3.1 Structural characterization of Spider Roll and mutant scaffold...............83

S3.2 Correlation between the predicted buried surface area and Rosetta binding energy for 43 native complexes......................84

S3.3 2D-[^1H, ^15N]-HSQC spectra of Spider Roll I58A and backbone resonance assignment.......................................................85

S3.4 Comparison of secondary structure determined by chemical shift index........................................................................86

S3.5 CS-rosetta structure prediction of Spider Roll I58A in an unbound state.................................................................86

S3.6 2D-[^1H, ^15N]-HSQC spectra of Spider Roll with PAK1 or PAK1 L470E........................................................................87

4.1 Selected ‘target’ proteins for Zn-mediated de novo interface design........102
4.2 Tetrahedral geometry parameters for Zn-coordination .................. 104

4.3 Rigid-body conformational search for Zn-mediated interface between ‘target’ and ‘scaffold’ proteins. ........................................... 105

4.4 Computational model of Zn-mediated interaction between Ubiquitin and a designed ‘scaffold’ .................................................. 106

4.5 Isothermal titration calorimetry (ITC) assays to show Zn binding to Spelter ................................................................. 108

4.6 Circular dichroism signal at 222 nm as a function of temperature in the presence and absence of Zn ........................................ 109

S4.1 Far-UV CD scan of Spelter and mutant Spelter as a function of Zn concentration ................................................................. 116
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AID</td>
<td>Auto-inhibitory domain</td>
</tr>
<tr>
<td>BME</td>
<td>2-mercaptoethanol or β-mercaptoethanol</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism</td>
</tr>
<tr>
<td>CFP</td>
<td>Cyan fluorescent protein</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothriptol</td>
</tr>
<tr>
<td>DMD</td>
<td>Discrete molecular dynamics</td>
</tr>
<tr>
<td>DDMI</td>
<td>Dock design minimize interface protocol</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>FP</td>
<td>Fluorescence polarization</td>
</tr>
<tr>
<td>ITC</td>
<td>Isothermal titration calorimetry</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>HSQC</td>
<td>Heteronuclear single quantum coherence</td>
</tr>
<tr>
<td>HYP</td>
<td>Hyperplastic discs protein</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>MBP</td>
<td>Maltose binding protein</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>MOPS</td>
<td>3-(N-morpholino)propanesulfonic acid</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>NOESY</td>
<td>Nuclear overhauser effect spectroscopy</td>
</tr>
<tr>
<td>PAK1</td>
<td>p21-activated kinase 1</td>
</tr>
<tr>
<td>PAK1-fl</td>
<td>Full length PAK1</td>
</tr>
<tr>
<td>PAK5</td>
<td>p21-activated kinase 5</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>REU</td>
<td>Rosetta energy units</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root mean square deviation</td>
</tr>
<tr>
<td>SASA</td>
<td>Solvent accessible surface area</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>S/N</td>
<td>Signal to noise ratio</td>
</tr>
<tr>
<td>TCEP</td>
<td>Tris(2-carboxyethyl)phosphine</td>
</tr>
<tr>
<td>TMAO</td>
<td>Trimethylamine N-oxide</td>
</tr>
<tr>
<td>Tris</td>
<td>2-amino-2-hydroxymethyl-propane-1,3-diol</td>
</tr>
<tr>
<td>UV</td>
<td>Ultra violet</td>
</tr>
<tr>
<td>WT</td>
<td>Wild-type</td>
</tr>
<tr>
<td>WT_IS</td>
<td>Wild-type inhibitory switch (truncated AID)</td>
</tr>
<tr>
<td>YFP</td>
<td>Yellow fluorescent protein</td>
</tr>
<tr>
<td>Zn</td>
<td>Zinc</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Recent development in computational protein design

The major challenge of computational protein design is to identify a most compatible sequence from a huge sequence space that can adopt a given target structure. The size of the sequence space is astronomical, for example a 50-residue protein can have as many as $10^{65}$ possible sequences. In order to identify ‘a needle in a haystack’ an efficient search algorithm and near-to-accurate energy functions are needed. In the last few decades, efforts from several groups have resulted in tremendous progress in this field.\textsuperscript{1, 2, 3} The most striking results have been in the \textit{de novo} design of known and unknown protein folds. Using a fixed backbone approach, Dahiyat \textit{et al} were able to design a completely new sequence for a zinc finger fold. The new sequence did not need a zinc ion for tertiary structure and NMR spectroscopy showed a compact well-ordered structure consistent with the designed model.\textsuperscript{4} Other pioneering work by Harbury \textit{et al} showed a successful incorporation of backbone flexibility for a large number of fixed amino acid sequences in the design of $\alpha$-helical bundle proteins with right hand superhelical twist.\textsuperscript{5} Kuhlman \textit{et al} went a step further by iterating between sequence optimization for a fixed backbone and backbone optimization for a fixed sequence to generate a globular protein with a novel fold.\textsuperscript{6} Though these successes reveal that computational methods can make it possible to search the huge sequence space for the protein design, a relatively unaddressed field of protein design is to design functional
proteins from scratch. A functional protein might show a novel interaction with an identified ‘target’ protein or might catalyze a new reaction. Recent achievements in designing novel enzyme catalysis show that the problem has been only lightly touched because the new enzymes have 5-7 orders of magnitude lower catalytic efficiency than natural enzymes.\textsuperscript{7; 8} Another unsolved problem is \textit{de novo} design of protein-protein interfaces. A novel interaction can help design biological tools and protein therapeutics by combining the enormous binding affinity and specificity of protein interfaces with rational choice of protein functionality.

**Computational protein design for functional improvements**

Published approaches of for improving the function of a protein have been restricted to searching small ranges of conformational space through mutagenesis while keeping the backbone atoms fixed. This approach, often called an inverse fold-problem\textsuperscript{9} considerably reduces the sequence-structure search space. Several works on specificity and affinity modulation have shown impressive results. Using a multi-state design approach to maximize the difference in free energy for a target conformation from the undesired competitor conformations, Havranek \textit{et al} was able to design homodimeric or heterodimeric coiled-coils with novel specificity motifs.\textsuperscript{10} Joachimiak \textit{et al} used rigid body transformation to sample ensembles of conformations of E7 DNase-Im7 immunity protein and then redesign the interface for lower binding energy. Using this approach the authors designed a novel hydrogen bonding network with a 300-fold specificity increase for designed pair over the design-WT pair.\textsuperscript{11} Earlier the same group used another approach, where a disruptive mutation was made on one side of an interface and then compensating mutations done on the other side, to design a novel specificity pair in the same system.\textsuperscript{12} Lippow \textit{et al} used a
computational procedure for optimization of charge-charge interactions with a single mutation and then combined multiple mutants for improved binding affinity of antibodies against the respective antigens. More recently Potapov et al showed that a complete module from a protein interface (a protein interface can typically consist of multiple modules, the residues across the module do not communicate) can be grafted into an unrelated protein to create an interface with similar affinity as the wild-type but a remarkably specific mode of binding. Reynolds et al did a computational redesign of the SHV-1 beta lactamase/beta lactamase inhibitor protein pair to improve the affinity by over 100-fold. Shifman et al made eight-fold calmodulin mutant that retained the affinity against one natural target while weakened affinity against other wild-type target peptides. A similar approach was used to improve the specificity of calmodulin by 900-fold for one target over another competitor. More recent work by Sammond et al showed that specificity switches can be made by creating knob-in-to-hole design and by changing charge-charge interactions to hydrophobic interactions. In all the above successful works, the similarity was in the approach of keeping backbone atoms fixed and considering only a small number of mutations during the redesign so that the backbone remained unperturbed. Designing a new protein interface in the regime of flexible backbone was left unexplored in these previous experiments.

There is also a high demand for improving the stability of proteins. Most popular ways to achieve the goal have been to improve the interactions in the core of the proteins. For example, using a rapid computational approach to search for stabilizing mutations, three mutations were identified in the core of an enzyme that increased its melting temperature by 10°C and half life at 50°C by 30-fold. Using the same computational program, another enzyme was stabilized by more than 5 kcal/mol by introducing only four mutations in the
core of the protein.\textsuperscript{20} Another complimentary approach to stabilize proteins has been introducing new charge-charge interactions on the surface of the protein.\textsuperscript{21, 22} Gribenko et al specifically selected polar residues (\textgreater 50\% solvent accessibility) on the surface of two different enzymes and mutated them to positive, negative or neutral residues irrespective of the charge on the neighboring residues and evaluated the total energy of the system. With a small number of substitutions (<5\% of total residues), the authors were able to improve the stability of the enzyme by >9 °C in melting temperature.\textsuperscript{23} Both these approaches for improving the stability of a protein use the simplifying approximations of keeping the backbone atoms fixed and/or using a small subset of residues for mutations.

Proteins with high solubility are required for high resolution structural studies, formulation of protein therapeutics and biochemical characterization of proteins. Molecular dynamics approaches\textsuperscript{24} have been used to identify potential residues on a protein which could be mutated to less hydrophobic or polar residues\textsuperscript{25} hence reducing the aggregation problem.

\textbf{Previous works in de novo design of protein-protein interface}

Computational design of a novel interaction from scratch is still an outstanding challenge. The major problems in designing a protein-protein interface are frequent burial of polar groups at the interface without any hydrogen-bonding partner and the requirement of coupling docking with design while incorporating for backbone movements.\textsuperscript{2} A recent success in interface design was achieved by exploring template-backbone orientations that would maximize the interaction between the partners. Both the partners were sequence optimized.\textsuperscript{26} Using this approach a heterodimer (modest high-micromolar affinity) was made from a small monomeric domain from protein G. Though the result was promising, a more challenging approach would have been a design with one partner kept fixed in sequence.
Another more promising result came from grafting key residues from an existing interface into a new ‘scaffold’ and then selecting for a good surface complementarity with the target protein. A low nanomolar affinity was achieved with this approach but this method loses viability in the absence of known binding partners.

The problem of de novo interface design can be simplified by choosing two domains that do not interact with each other but do interact with structurally similar partners. The major problem of achieving high surface complementarity between the interfaces is solved in this case. Using this approach a chimeric homing endonuclease was created by choosing one domain each from distantly related endonucleases and then optimizing the interface for binding.

Contrary to the modeling-based approaches, directed evolution techniques have been very successful in designing novel protein binders. Novel lower nanomolar maltose binding protein binders were selected using phage display with FN3 domain as a starting protein or using ribosomal display technique with an ankyrin repeat based protein. These approaches though have caveats as there is only limited control over choosing the binding site. Computational approaches also have an advantage over directed evolution in cases where a multi specific binder needs to be designed with no affinity against some other ‘targets’.

**Rosetta molecular modeling suite**

The Rosetta molecular modeling software package is being developed simultaneously by several research groups. It was initially created for structure prediction and then subsequently used for protein docking, loop modeling, high resolution structure refinement and protein design. The structure prediction or protein folding problem has been adequately
addressed by Rosetta and it has been very successful in predicting tertiary structures of sequences less than 100 amino acids. The protein design or inverse folding problem, where a low energy sequence needs to be determined for a given structure, has been also successfully solved using Rosetta. The two inversely related problems both need a reasonably accurate energy function and a sampling method capable of finding the minima of this function.

*Energy function:* Rosetta’s full-atom energy function is dominated by 12-6 Lennard-Jones potential term. This energy term consists of a favorable attractive term ($E_{\text{attractive}}$) for atoms to be close to each other and an unfavorable repulsive term ($E_{\text{rep}}$) to penalize atoms clashing with each other. In order to compensate for the fixed backbone assumption during design, $E_{\text{rep}}$ is dampened to incorporate for some level of backbone flexibility. The other component of the energy function include a Lazaridis-Karplus implicit solvation term ($E_{\text{solv}}$) that favors the exposure of hydrophilic residues at the surface and penalizes for the burial of polar groups. An orientation dependent hydrogen bonding term (derived from four parameters (distance between the hydrogen and the acceptor atom, angle at the hydrogen atom, angle at the acceptor atom and the dihedral angle between given by rotation around the acceptor-acceptor base bond) between an acceptor and a donor atom allows for the burial of polar groups if they are able to make hydrogen bond with another polar group. The residue pair potential term ($E_{\text{pair}}$) in Rosetta is a knowledge-based term for scoring electrostatic interactions. Charge-charge interactions are very sensitive to the local environment and difficult to model. The $E_{\text{pair}}$ term is derived from the probability of finding two charged residues near each other in the Protein Data Bank (PDB). Another component of Rosetta energy function is the rotamer self energy term ($E_{\text{dun}}$). This term is also a knowledge-based
term and is derived from the probability of finding a side-chain in a particular conformation with certain phi-psi dihedrals of the backbone in PDB. 40 The backbone torsion potential ($E_{\text{rama}}$) term is also a knowledge-based term, and is related to Ramachandran torsion preferences. It is derived from PDB statistics by measuring the probabilities of seeing particular amino acids in a secondary structure type (helix, strand and loop) for a particular phi, psi angles. 41 From a weighted linear sum of the above energy terms, an unfolded state energy term ($E_{\text{ref}}$) is subtracted. $E_{\text{ref}}$ and the weights of each energy term are calculated to best reproduce native sequences for known structures. 42

Rosetta also uses coarse-grained centroid-based representation of side-chains when an excessive backbone sampling is required. Optionally it can be used in case of docking to get a right orientation of the two chains and relieve any backbone clashes and if it is followed by all-atom designs. Rosetta coarse grained energy function is dominated by an environment term ($E_{\text{env}}$) and pair term ($E_{\text{pair}}$) which respond to solvation and electrostatic effects and are based on observed residue distributions in protein structures. Hydrogen bonding is not explicitly described as most atoms are missing, but favorable scores are given for the formation of β-strands and sheets. Lennard-Jones repulsive term ($E_{\text{rep}}$) which penalizes for clashes between the backbone or centroid-based side chains is included while the attractive term is included in the form of a radius of gyration ($R_g$) term which rewards globally compact structures. 33

The approximate nature of the energy function is apparent with the absence of explicit solvent, long range electrostatics, residual dynamics in the molecule and conformational entropy. 34
**Search algorithm:** Rosetta uses the Metropolis Monte-Carlo search algorithm with simulated annealing to identify the local minimum of the energy function. It is a stochastic approach and useful in searching through a number of local minima. It is non-exhaustive and uses discrete set of side-chain conformations (rotamers). The protocol starts with a random initial conformation of the system followed by a perturbation such as the substitution of a discrete rotamer. Energy is calculated before and after the substitution is made. If the energy of the system drops, the change is accepted. If the energy of the system goes up, the change is accepted or rejected based on the Metropolis criterion. In brief, the Metropolis criterion allows occasional increase in energy so that the system is not trapped in a local minimum. For simulated annealing, the search is repeated at lower temperature. This guides the system to the local minimum because, at lower temperature, the system is less likely to ‘jump’ out of it. A trajectory may consist of thousands of substitutions which is typical for convergence.

Monte Carlo-plus-minimization strategy\(^{43}\) is followed for more rigorous rotamer optimization. After each Monte Carlo optimization, torsion angles of the backbone and side chains are perturbed from ideality to promote a gradient-based descent of the conformation to the local minimum.

**Affinity reagents for biosensors**

For our purpose, an affinity reagent is a protein that can bind any other given protein. An affinity reagent can be highly specific and can even distinguish between two different states of the ‘target’ protein (Figure 1.1). Affinity reagents have potential uses as therapeutics or for creating biological tools. A binder which can preferentially bind to the activated form of a ‘target’ protein can be an important tool for visualizing spatio-temporal dynamics of the protein inside living cells. An affinity reagent derived from the WASP domain\(^{44}\) and from the
Rho binding domain\textsuperscript{45} have been used to create biosensors against the activated form of CDC42 and RhoA respectively. These biosensors were very useful in visualizing the timing and localization of the above Rho GTPases and understanding their implications on cell motility and polarization under different environmental cues.

An affinity reagent for a biosensor needs a signal producing moiety which can either be genetically encoded (RhoA biosensor\textsuperscript{45}) or can be exogenously attached as a solvent-sensitive dye (CDC42 biosensor\textsuperscript{44}).

![Figure 1.1. Proposed function of an affinity reagent.](image)

(a) The auto-inhibited state of the target protein (functional domain in ‘yellow’ and auto-inhibitory domain in ‘green’). A biosensor consisting of an affinity reagent is shown in ‘cyan’ and a solvent-sensitive dye is shown as blue star. (b) Activation of the target protein results in conformational change and dislodging of the auto-inhibitory domain. When the biosensor interacts with the active protein, it partially desolvates the dye leading to fluorescence intensity change.

**Model system, goal and design approach**

A good model system against which an affinity reagent is to be designed should show very distinct conformations in the activated and inactive forms. An affinity reagent designed against the active conformation will then have a very minimal interaction with the inactive form and will have the potential to be converted into a biosensor. A good model system should also be an important functional protein so that a biological tool against it has more impact. A model system for computational design should also have structural information, at least one good quality crystal structure.
p21-activated kinase 1 (PAK1) is a serine/threonine kinase that undergoes an enormous conformational change when an activator like CDC42/Rac binds to it. When an activator binds to the auto-inhibited dimer of PAK1, it results in destabilization of the folded structure of the auto-inhibitory domain (AID), which is then followed by stabilization of activation loop via phosphorylation of T423. Subsequent phosphorylation on the AID (e.g. S144), prevents any simple reversal of activation state (Figure 1.2).

![Figure 1.2. Activation mechanism of p21-activated kinase 1 (PAK1).](image)

PAK1 is involved in multiple functions inside the cell. It participates in cell motility and cytoskeletal dynamics, transcription though MAP kinase cascades, death and survival signaling and cell cycle progression. Consequently, dysregulation in PAK1 functions have implications in pathological conditions and cell transformation. With so many functions to perform, we hypothesize that the localization and timing of activated form of PAK1 must be tightly controlled. Hence, a biosensor against activates PAK1 will be an important tool to offer insights into the already existing knowledge.
Crystal structures of autoinhibited PAK1 (PDB ID: 1F3M)\textsuperscript{46} show the relative orientation of the kinase domain and the AID (Figure 1.3). The whole structure consists of two kinase domains and two auto-inhibitory domains. A $\beta$-strand swap between the AIDs results in a dimerization interface and has been referred to as the CRIB (for CDC42/Rac interactive binding region). The activators CDC42/Rac have been proposed to bind at that site. Crystal structures of the PAK1 kinase domain, with and without phosphorylation mimic in the activation loop also exist (PDB ID: 1YHV, 1YHW).\textsuperscript{48} The two crystal structures have similar conformation, which suggest that phosphorylation of the kinase domain may not be necessary for attaining the active conformation, though it might be required for activity. The information from these structures will be useful for computational modeling.

![Figure 1.3. Crystal structure of auto-inhibited PAK1. The structure (PDB ID: 1F3M) shows two kinase domains (chains C and D) and two AIDs (chains A and B).](image)

The goal of my research is to create an affinity reagent for the activated form of PAK1. The cleft formed by helices (\(\alpha\)EF and \(\alpha\)G) is left vacant when PAK1 gets activated and is a promising region for protein-protein interaction. I used the PAK1 kinase domain
from the structure of inactive PAK1 (1F3M, Chain C) as a model for the active PAK1. When the crystal structure of kinase domain alone (1YHW) is aligned with the kinase domain from the structure of inactive PAK1 (1F3M, chain C), a low backbone RMSD of 1.9 Å is observed. Two major differences can be highlighted (Figure 1.4). First, there is a difference in relative orientation of the cleft (helices αEF and αG) where AID binds. The structure of kinase domain alone, 1YHW, shows the helices are more open. Another difference is seen in the conformation of activation loop. 1YHW shows an ordered structure, in 1F3M, the loop is disordered. It may be possible that activation loop requires an anchoring to the residue V385 near the active site. In 1F3M, the inhibitory tail extends to the active site and anchors to the residue V385.

Figure 1.4. Structural alignment of PAK1 kinase domain and kinase domain from auto-inhibited PAK1. The crystal structure of the PAK1 kinase domain alone (PDB ID: 1YHW, ‘magenta’) is very similar to the activated kinase domain, but there are two major differences from the kinase domain of auto-inhibited PAK1 (PDB ID: 1F3M, ‘yellow’), the absence of activation loop in the inhibited kinase domain and rearrangement of αEF and αG helices.
In order to design a binder of PAK1 kinase domain, we decided to use the kinase domain from 1F3M that showed the conformation of protein in a bound state. The idea was that a binder to the cleft in the kinase domain (1YHW) would induce a conformational change and bring the helices (αEF and αG) closer to each other which would be consistent with 1F3M. Two different approaches were used to design an affinity reagent for PAK1 kinase domain (Figure 1.5). In one, a naturally occurring binder can be redesigned for improved stability, specificity, solubility and affinity. The other approach was de novo design of a binding interface on another stable protein (‘scaffold’).

**Figure 1.5. The two approaches for designing affinity reagents.** (a) Redesigning a natural binder for improved function. The functional improvement can be in terms of affinity, specificity, solubility or stability. (b) De novo interface design approach where a ‘scaffold’ protein is redesigned for a novel interaction with a ‘target’ protein.

The AID binds to the kinase domain in the auto-inhibited state of PAK1. This is the binding interaction we will mimic for our biosensor. Under biological condition, an activator (CDC42/Rac) binds to the AID and releases it from the kinase domain cleft. The inhibitory tail of the AID that extends to the active site and contact with V385 is released and the activation loop becomes ordered. If the AID is truncated to remove the inhibitory tail, and
residues involved in interaction with the activator removed, the minimal domain should still be a good binder of the kinase domain. Using a computational approach in Rosetta, the minimal domain was redesigned to improve the features of a good affinity reagent that include solubility, specificity, stability and binding affinity. The work has been described in chapter 2.

Another approach to this problem would be to find a stably-folded protein that is structurally similar to the AID, and use it as a seed of an affinity reagent. This similarity is necessary because a major concern of de novo interface design is surface complementarity. A segment from the hyperplastic discs protein (PDB ID: 1I2T) shows a good structural homology with the AID and was used as ‘scaffold’ protein for interface design. We developed a new protocol for protein interface design in Rosetta. The protocol was then utilized to create an interface on the ‘scaffold’ protein resulting in a modest binder. The work has been described in chapter 3.

**Incorporating metal ions in the protein interface as a future direction**

*De novo* design of protein-protein interaction success can be improved to some extent by starting with shape complementarity between the ‘scaffold’ and the ‘target’. The ‘scaffold’ protein can have good overall structural homology with a native binder of the ‘target’ or may be the region on the ‘scaffold’ to be designed has a curvature complementary with the ‘target’ surface. The former approach has been used in this work (described in chapter 3). The later approach was used in cases of successful published *de novo* designs. Another approach would be to design specific interacting motifs on the two sides of the interface. This can be achieved by designing a ‘knob’ on the ‘scaffold’ protein which is complementary to a ‘hole’ on the target or vice-versa. Modulating long range interactions have also strong effects.
on association rates. Patches of opposite charge groups on the two sides of interface can also play an important role in improving association rates. The charged groups can even be placed in the vicinity of the actual protein interface and still significantly affect the association rate. Another approach would be to create a metal-mediated interaction between the interfaces. Apart from contributing to the enthalpy of binding and affecting the dissociation rates, metal-mediated interactions have also shown to affect the association rate by influencing the alignment of two partners.

We used this approach to create a metal (zinc) mediated interaction between ‘target’ protein UbC12 (an E2 enzyme in the neddylation pathway) or Ubiquitin and several other ‘scaffold’ proteins chosen from the protein data bank. UbC12 and Ubiquitin have an exposed histidine surrounded by some hydrophobic residues. Since zinc has a tetrahedral coordination geometry, where four ligand atoms coordinate with it, the idea was to design three residues (consisting of His or Cys) on the ‘scaffolds’ and then the fourth coordination contributed by the preexisting ‘target’ histidine.

This work discussed in chapter 4, shows partial success in computational design. We succeeded in designing a metal binder, but the metal mediated-interaction of two proteins has still not been successful.
References


CHAPTER 2
COMPUTATIONAL REDESIGN OF PAK1 AUTO-INHIBITORY DOMAIN FOR A BIOSENSOR

Abstract
Computational redesign of the auto-inhibitory domain (AID) of p21-activated kinase 1 (PAK1) has been described. AID is a 72 residue domain (residues 78-149 of PAK1) remains bound to kinase domain of PAK1 in auto-inhibited state. AID-kinase interaction can be split onto two distinct modules, first domain-domain interaction between inhibitory switch (residues 87-136 of PAK1) and helices αEf and αG of the kinase domain and the second peptide-domain interaction between peptide (residues 144-146) and kinase domain active site region (residues 385-387). Multistage activation process of PAK1 results in dislodging of inhibitory switch from the kinase, resulting in exposed αEf and αG helices. An affinity reagent can specially be targeted to the exposed residues on the helices as they contribute to a ‘hotspot’ region for PPI. Truncated AID (or WT_IS, residues 83-137 of PAK1) was redesigned to have improved properties for a good affinity reagent that can bind the ‘hotspot’ region and can be developed into a biosensor to detect the activated form of PAK1. The binding interface of AID to PAK1 was retained during the redesign process. Three generations of designs were tried. In the first case, a β-hairpin was designed, in the second case a 20-residue helical region was created for a new N-terminus and lastly 16-residue helix was created at the C-terminus of truncated AID. The goal was to improve the structural
property, solubility and binding affinity (for the kinase domain, our model for activated form of PAK1) of the designed domain. The first design (Hairpin) did not show any advantage over truncated AID. The second design (AlmostHelix) showed gain in structure and solubility. The final design (PAcKer) when inserted into CFP showed ten-fold improvement in binding affinity for the kinase domain. In all designs an extra secondary structure region was created in an attempt to improve the functional property. The work is novel in that way as no example in literature shows that a new secondary structure sequence can be attached to an already existing sequence for improvement in function.
**Introduction**

Natural proteins are optimized for protein-protein interactions under certain conditions. Most often the proteins are evolved to have multiple interactions in order to regulate a signaling pathway. Sometimes the proteins can be poorly structured in the absence of the binding partner and more often the solubility can impose limitations when the protein is expressed all by itself. These possibilities limit the usage of a native protein as an affinity reagent. Our ability to rationally design these proteins for improved binding affinity, specificity, stability or solubility can result in superior candidates for affinity reagents. In past decade, both computational and experimental techniques have resulted in a number of functional improvements in proteins. Most of the works carried a series of mutagenesis while hypothesizing the structure remained unperturbed. A recent work showed that an extended sequence from a structural homologue can be used in a functional protein without significantly perturbing the beta-lactamase activity.\(^1\) Other *de novo* design work of a loop region in a beta-sandwich resulted in an improved stability of the protein.\(^2\) Some directed evolution works with a FN3-type scaffold has resulted in a novel BC or FG loop with a novel binding affinity for a ‘target’ protein.\(^3\) A largely unaddressed protein engineering field has been *de novo* design of a secondary structure region of a protein that would result in a functional advantage.

The auto-inhibitory domain (AID) of p21-activated kinase domain 1 (PAK1, PDB ID: 1F3M) is a dimer with the formation of a β-ribbon at the interface (Di). The multistage activation process of PAK1 starts with a binding of activator to the Di, resulting in a partial
unfolding of the AID and dislodging from the kinase domain.\textsuperscript{4} A series of auto-phosphorylation at multiple serine sites result in stabilization of the ‘open’ state. AID thus has specificity for multiple proteins that include the kinase domain of PAK1 and the activators, Cdc42 and Rac. The published structures of Cdc42 or Rac with CRIB (CDC42/Rac-interactive binding domain) of PAKs\textsuperscript{5; 6; 7; 8} show the involvement of Di segment of AID. It also has a propensity to dimerize and undergo structural changes when activator protein binds to it.

In order to use AID as a specific affinity reagent to be developed for biosensor, it needs to be redesigned so that it is monomeric, structured, stable and if possible has improved binding affinity. AID all alone has partially exposed core residues that reduce the solubility. An improvement in solubility will have an added advantage for an affinity reagent. The affinity reagents derived from natural binders have been used to create biosensors against Cdc42\textsuperscript{9} or RhoA\textsuperscript{10} and were used to visualize spatio-temporal dynamics of the activated form of respective protein.

A recent successful PAK1 biosensor was created by fusing YFP and CFP at the two termini of PAK1.\textsuperscript{11} The biosensor was effective in visualizing the PAK1 activation during spreading and motility. Useful information regarding the activation was also gathered. But the usefulness of modifying proteins for detections is limited. Modification of the proteins of interest for a biosensor can result in perturbed function. Usually the modified proteins also need to be overexpressed beyond the endogenous level. More than that the PAK1 construct did not have the binding sites of Nck and Grb2 and hence adapter dependent effects could not be observed. A more efficient biosensor would sense the endogenous level of activated protein.
A solvent-sensitive dye based biosensor for Cdc42\textsuperscript{9} and a photoactivable biosensor for Rac\textsuperscript{12} were created earlier. Cdc42 and Rac1 are the activators of PAK1. The biosensor of Cdc42 revealed microtubule-dependant activation of the protein in the cell periphery and not in the filopodia, activation of Cdc42 in the trans-Golgi compartment and a tight coordination between cellular extension and retraction accompanied by the activation of the protein. In another case, genetically encoded photoactivatable Rac1 showed cell protrusion and ruffling when shone with blue light. PAK1 is also an immediate effector protein of Rac1 but Rac1 is involved in activation of several other proteins including PAK2-4. A biosensor for PAK1 can track the spatial and temporal dynamics of the protein specifically.

In this work we redesigned AID of PAK1 as an independent monomeric protein so that it can be used as an affinity reagent. The goal was to retain its kinase binding property, remove its inhibitory property and disrupt its affinity for the activators of PAK1. In three generations of design a secondary structure sequence was created in an effort to improve the functional property of AID as an affinity reagent.

Results

AID is not a good affinity reagent for a biosensor

AID has an inhibitory tail (residues 138-149 of PAK1) that goes into the active site of PAK1. The dimerization segment (residues 81-87) is involved in a β-strand swap with other AID monomer. In our preliminary experiment to use WT_IS (residues 83-137 of PAK1) alone as an affinity reagent for PAK1, we found that the protein did not show strong helical content (Figure 2.2) and hence would be unfolded. We hypothesized that the two anchoring events (binding of inhibitory tail to the active site and the β-strand swap) along with binding to the kinase domain were important to achieve the folded nature of AID as described in crystal
structure (PDB ID: 1F3M). Since WT_IS was not properly folded, the exposure of hydrophobic core resulted in low solubility at the purification stage and also made it susceptible to proteolysis. In an ITC experiment, WT_IS bound to kinase domain with a modest affinity of 4 µM. Hence, we decided to redesign the AID for improved structure, stability, solubility and binding affinity.

**Design strategy**

We created 3 generations of designs using AID as the starting structure (Figure 2.1 a). In all these designs, the goal was to keep the kinase-interacting regions intact (residues 112-134) while redesigning away from the interface for improved functional properties. A multiple sequence alignment shows that majority of the residues in the range 112-134 that was facing the kinase domain were left invariant while some mutations were made on the helix side facing away from the kinase to accommodate improved packing with the designed sequences (Figure 2.1 b).

The focus of redesign process was to improve the core of the WT_IS (residues 83-137 of PAK1). As mentioned earlier, in the absence of the inhibitory tail (residues 138-149), WT_IS did not fold properly. A β-strand swap between two auto-inhibitory domains may be also responsible for stabilizing the folded conformation of the AID but results in dimerization. Hence, in our first approach, we built a β-hairpin between the residue 85 of chain B and residue 83 of chain A (PDB ID: 1F3M) to join the two β-strands. Since β-ribbon formation was intra- instead of inter- domain, it would inhibit the β-strand swap event necessary for dimer formation. Residues 83 and 86 (conserved His across effector proteins of CDC42 and Rac) interact with a conserved Asp on the activator protein is a critical
interaction for the activation process\textsuperscript{5,13}. Hence, we removed the two His in our designs. The design was named Hairpin.

**Figure 2.1. Structural and primary sequence comparison of the models evolved from the auto-inhibitory domain (AID, cyan).** The focus was to design sequences away from the PAK1 kinase domain (yellow surface) interaction region. A β-strand (magenta) is contributed by another AID. (a) Three generations of designs are shown as ‘Hairpin’, ‘AlmostHelix’ and ‘PAcKer’. The designed structural regions are shown in green. Cyan regions were left largely intact except for some point mutations to accommodate the new secondary structure regions. (b) Multiple sequence alignment shows the designed secondary structure regions were at N-terminus or C-terminus while the residues involved in binding to the kinase domain were left invariant.

In a second approach, a 20-residue helical stretch was built in the N-terminus of trunc-WT\_IS-1 (residues 100-137). The new region in the model packed against the other helices of trunc-WT\_IS-1 and hence gave a tighter core. Since the new region did not have any propensity of forming a β-strand, possibility of inter-domain swap was negligible.
Potential residues which interact with the activators of PAK1 were also removed and the partially exposed core residues were buried too. The model was called AlmostHelix.

In a third approach a 16-residue helical stretch was created at the C-terminus of the trunc-WT_IS-2 (residues 87-137). The goal of this approach was to bring the N-terminus and C-terminus of the domain close to each other so that it can be inserted into a fluorescent protein (for example CFP) between the residues Glu-172 and Asp-173. The approach of using a fluorescent protein with a biosensor has been proposed earlier for ratiometric imaging to correct for varying cell thickness and non-uniform illumination of the cell.14 The new helical stretch was also intended to bury the core residues for improved solubility and stability. Most of the regions of WT_IS involved in CDC42/Rac interaction or domain swap were also removed in this design. The model was called PAcKer (CFP-PAcKer when inserted into CFP).

**Design protocol**

We used an iterative structure and sequence optimization protocol for loop modeling within the framework of Rosetta software package.15, 16 In brief the protocol uses three- and nine-residue fragments drawn from the database to build loops of appropriate size. The fragments were picked for certain secondary structure propensity. For example, poly-alanine sequence was used to collect helical loops from the database and a hypothetical sequence TITNGTIT for collecting loops with high β-hairpin propensity. The loop was then redesigned for best sequence-backbone compatibility as discussed in Methods section. Distance constraints were issued for a proper loop orientation with respect to the rest of the protein.
**Expression and solubility of the designs**

WT_IS, Hairpin and PAcKer were expressed with an MBP tag while AlmostHelix was expressed with a GST tag in *E. coli*. PAcKer inserted into CFP (called CFP-PAcKer) was also expressed with a 6×His tag. AlmostHelix expression level in soluble fraction was distinctly higher than the other proteins and it was clear from later observations that it was more due to the AlmostHelix sequence and not due to vector. Once the protein was cleaved of the GST or MBP fusion, and concentrated, AlmostHelix showed no signs of aggregation up to a level of 1 mM. WT_IS and Hairpin had much lower solubility (< 200 µM) while PAcKer showed an intermediate level of solubility.

**Structural properties of the designs**

Far-UV circular dichroism (CD) scan showed WT_IS, Hairpin and PAcKer unfolded while AlmostHelix showed a good helical content (minima at 208 nm and 222 nm, Figure 2.2 a). Absence of any structure in WT_IS also confirmed that the structural change might be induced by the two interactions in which AID is involved, first being the binding to the active site of kinase domain using inhibitory tail and the second being the formation of β-ribbon using a domain swap with another AID.

Thermal denaturation of the designs at 222 nm showed AlmostHelix having modest cooperative unfolding. This confirmed that the protein is folded but lack of sharp transition during thermal denaturation also raised the possibility of a molten globule which would suggest a very low difference in total energy between the folded and unfolded states. Hence, we performed a far-UV CD scan at a high concentration of trimethylamine N-oxide (TMAO), an osmolyte which has been shown earlier to have an ability to force thermodynamically unfolded proteins to fold. We saw a considerable increase in helical
content at 0.5 M and 1.0 M of the osmolyte (Supplementary figure S2.1 a). When thermal denaturation of AlmostHelix was carried at 222 nm in the presence of 2 M TMAO, we found a steeper unfolding transition (Supplementary figure S2.1 b). This confirmed that osmolyte can increase the difference in energy between unfolded and folded states and populate the folded state.

**Figure 2.2. Structural characterization of computational models and WT domain using CD.** (a) AlmostHelix shows good helical content. WT_IS, Hairpin and PAcKer are all unfolded. (b) AlmostHelix shows very modest cooperative thermal denaturation.
Binding affinity and suitability as an affinity reagent

One of the most important features of an affinity reagent is its binding affinity. A high binding affinity ensures improved sensitivity when the affinity reagent is used as a biosensor. WT_IS and the designs were tested against PAK1 kinase domain (our model for active conformation of PAK1) and PAK1 full length (our model for inactive PAK1). Two mutations, V127E and S144E on PAK1 full length ensured AID in an unbound state. This mutant PAK1 was a model for ‘open’ form of PAK1. Isothermal titration calorimetry experiments showed CFP-PAcKer bound to the kinase domain of PAK1 with 10-fold tighter binding affinity (gain in binding energy by ~1.4 kcal/mol) than the WT_IS (Table 2.1). We hypothesize that the 16-residue rigid linker created on trunc-AID-2 and fusing the new N-terminus and C-terminus to the CFP reduces the loss in entropy upon binding. Earlier the CD scan had confirmed WT_IS to be unfolded and raised the possibility of binding induced folding.

Table 2.1. Biophysical properties of the designs and comparison with the truncated AID. The binding affinity was measured for PAK1 kinase domain (model for ‘active’ conformation). WT_IS corresponds to truncated AID of PAK1 (residues 83-137).

<table>
<thead>
<tr>
<th>Structured(^a)</th>
<th>WT_IS</th>
<th>Hairpin</th>
<th>AlmostHelix</th>
<th>PAcKer</th>
<th>CFP-PAcKer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal melt(^b)</td>
<td>Non-cooperative</td>
<td>Non-cooperative</td>
<td>Cooperative</td>
<td>Non-cooperative</td>
<td>ND</td>
</tr>
<tr>
<td>(K_d) (µM)(^c)</td>
<td>4</td>
<td>2.6</td>
<td>1.6</td>
<td>ND</td>
<td>0.4</td>
</tr>
</tbody>
</table>

\(^a\): Far-UV CD scan  
\(^b\): CD signal at 222 nm as a function of temperature  
\(^c\): Isothermal titration calorimetry. Fluorescence polarization assay (FP) was used only for AlmostHelix
This advantage seemed to be absent in AlmostHelix. Though the domain appeared to be more structured than WT-IS, there was no advantage in terms of binding energy. One possible explanation for this can be made in terms of ‘locking’ of the domain in unsuitable orientation resulting in loss of conformational sampling.

CFP-PAcKer bound to PAK1 kinase domain with an affinity of ~400 nM (Figure 2.3 a). The affinity reagent failed to bind to PAK1 full length (Figure 2.3 b). The binding affinity against PAK1^{V127E/S144E} (‘open’ variant) was around 10-fold weaker (Supplementary figure S2.2).

Figure 2.3. Isothermal titration calorimetry assays to determine the binding affinities of CFP-PAcKer against active and auto-inhibited conformations of PAK1. (a) 208 μM of PAK1 kinase domain (model for active conformation) was titrated against 18 μM CFP-PAcKer. An equilibrium dissociation constant of 400 μM was observed with a molar ratio ~1. (b) 262 μM PAK1 full length (model for ‘close’ auto-inhibited form) was titrated against 21 μM CFP-PAcKer. The titration failed to show any appreciable heat released or absorbed. The data sets were drawn on similar scale for comparison.
In order to find out that the binding of CFP-PAcKer to the kinase domain is consistent with the hypothesis, a mutation V→E corresponding to residue 127 (based on PDB ID: 1F3M) was made on PAck. Interaction of the CFP-PAcKer\textsuperscript{V127E} was completely disrupted. Similarly a mutation L→E corresponding to residue 470 on PAK1 kinase domain (based on PDB ID: 1F3M), also abolished the binding between the two proteins (Supplementary figure S2.3).

**CFP-PAcKer derived biosensor (mero-CFP-Packer) specifically ‘lights up’ in the presence of PAK1 kinase domain in vitro**

A mutation K→C was made at the position corresponding to residue 134 (based on PDB ID: 1F3M) on CFP-PAcKer. A solvent-sensitive merocyanine dye\textsuperscript{17} was conjugated at this position. Since this position was at the interface, a big change in dye environment was expected upon PAck-PAcKer interaction resulting in alteration of the fluorescence property of the dye. The dye has an excitation wavelength of 593 nm and emission wavelength at 620 nm. The fluorescence property of CFP (excitation at 433 nm and emission at 474 nm) was used as a background signal and used for ratiometric expression. The normalized ratio \(((\text{Dye/CFP})_{\text{Tritration}} / (\text{Dye/CFP})_{\text{Buffer}})\) increased during the titration of PAK1 kinase domain. A ‘one-site binding’ fit gave a binding affinity of around 600 nM, consistent with the equilibrium dissociation constant observed in ITC. Biosensor did not show change in normalized fluorescence ratio with full length PAK1 or PAK1 kinase\textsuperscript{K40E} mutant (Figure 2.4). When dye was conjugated to CFP-PAcKer\textsuperscript{V127E} there was no change in ratiometric fluorescence with PAK1 kinase domain titration (Supplementary figure S2.4).
Figure 2.4. Biosensor derived from CFP-PAcKer selectively ‘lights up’ in the presence of PAK1 kinase domain. A net gain of ~40% in the normalized fluorescence ratio was observed in case of PAK1 kinase domain when it was titrated in 200 nM of biosensor. PAK1 kinase$^{L470E}$ and PAK1 full length failed to show any appreciable change in fluorescence ratio. PAK1$^{V127E/S144E}$ showed an intermediate level of change in fluorescence ratio. A decrease in signal at higher titrations may be attributed to prolonged incubation.

Specificity of CFP-PAcKer derived biosensor

CFP-PAcKer derived biosensor is expected to ‘light up’ against group 1 PAKs (PAK 1-3). Since group 2 PAKs (PAK 4-6) are more divergent based on difference in domain architecture, we tested CFP-PAcKer against PAK5 kinase domain expressed with a GST tag. Isothermal titration calorimetry did not show any appreciable heat absorbed or released (Supplementary figure S2.5). CFP-PAcKer conjugated to merocyanine fluorescent dye showed very low change in ratiometric fluorescence in the presence of 1 µM of GST-PAK5 kinase domain (Figure 2.5). A full titration up to 10 µM of GST-PAK5 showed a saturation level for the normalized fluorescence different from the saturation level observed for PAK1 kinase domain (Supplementary figure S2.6). This was not very surprising because the
fluorescence signal depends on two different parameters, first how tight the biosensor is binding to the ‘target’ and second how efficiently the dye gets desolvated. The helices αEF and αG where the biosensor was supposed to bind, is more polar in PAK5 than in PAK1 and a lower affinity for the biosensor was expected. Different residue composition on the PAK5 kinase domain interface might result in a different mode of biosensor interaction resulting in higher desolvation of the dye environment and a very different fluorescence property of the dye. Over all the results confirmed that CFP-PAcKer based biosensor was more selective to group 1 PAKs and contributions in signal could be expected from group 2 PAKs if the localized concentration reached at micro-molar levels.

Figure 2.5. CFP-PAcKer derived biosensor gives a poor response with PAK5 kinase domain. GST-PAK5 kinase showed a minimal change in normalized fluorescence ratio at 1 µM of titrant. For a proper comparison, GST at 1 µM or equimolar (1 µM) PAK1 kinase domain and GST were also titrated in biosensor. ‘Blank’ refers to 200 nM of biosensor alone.
Discussions

We redesigned the auto-inhibitory domain of PAK1 for improved function as an affinity reagent. The focus of the design process was to build the core of that domain so that it can attain a tertiary structure. A well-defined structure can have advantage in terms of binding affinity (reducing the cost of conformational entropy loss), stability (longer shelf-life), solubility (improvement due to less exposed hydrophobics) and resistance to proteolysis. Only in case of AlmostHelix, we saw improvement in stability and solubility but that did not show any advantage in binding affinity.

A 16-residue rigid-linker created in PAcKer that helped to bring C-terminus close to N-terminus allowed an efficient insertion of the domain into the CFP. CFP was required in the final biosensor for ratiometric expression of signal from the dye. The distance between N- and C-terminus of folded minimal AID (WT_IS-2, residues 87-137) is ~22 Å. To insert the domain between E172 and D173 would require an ample amount of optimization with different forms of linkers. For example just using Gly-Ser (GS) rich linkers, the challenge will be to come up with an adequate number of repeats so that the domain does not come too close to CFP or get too much strained upon binding to the kinase domain. Very long linkers will make the domain very floppy with respect to CFP. A construct (CFP-IS) linking E172 of CFP with N-terminus of WT_IS-2 with 3×GS and C-terminus of WT_IS-2 with D173 using 2×GS did not show any advantage in terms of binding affinity (Supplementary figure S2.7). Computational methods helped to design an adequate length of linker in C-terminus of the PAcKer. When PAcKer was inserted into CFP, binding induced folding was assisted in the same as we hypothesized with the AID, where the two termini of AID were restricted in
space, N-terminus formed β-strand swap with another AID and C-terminus tail bound to the active site of the kinase domain.

It will be interesting to see CFP-PAcKer derived biosensor inside cells. Overall change in ratiometric signal with PAK1 kinase domain titration was 40%-50% which might still be enough to visualize PAK1 activation inside the cell. Several FRET-based biosensors have been shown to reach very similar level of signal change in bound state from unbound state.\textsuperscript{10, 11}

**Materials and methods**

**Loop modeling using Rosetta**

We used loop modeling protocol as described earlier.\textsuperscript{18} The protocol uses Rosetta scoring function and fragment insertion methodologies which have been utilized for \emph{de novo} structure prediction.\textsuperscript{15} Using Robetta server\textsuperscript{19}, a customized library of fragments was created from database of high resolution structure for each three- and nine-residue window in a fasta sequence. A pseudo-sequence for the region to be modeled was used to bias the fragments to a desired secondary structure. For Hairpin, the query fasta sequence consisted of fixed regions (Residues 79-85 Chain B, 88-137 Chain A of PDB 1F3M) and TITNGTIT in between the two chunks. For Almost helix the query fasta sequence was poly-alanine sequence (20 residues) followed by sequence consisting of residues 100-137 of AID (Chain A of PDB 1F3M). In the final case, the query fasta sequence for PAcKer was created using sequence 87-137 of AID (Chain A of PDB 1F3M) followed by a 16 residue poly-alanine sequence. The fragments corresponding to the region to be modeled were then assembled using Monte Carlo simulated annealing technique. Fragment insertions were accompanied by a “wobble” operation where the backbone torsion angles (φ, ψ) were perturbed. Monte Carlo
minimization protocol\textsuperscript{15, 20} was used to simultaneously optimize sequence for a fixed backbone and lowest energy backbone for a fixed sequence. All-atom energy function was optimized.\textsuperscript{15} In brief the function is a linear weighted sum of 12-6 Lennard-Jones potential, the Lazaridis-Karplus implicit salvation model\textsuperscript{21}, an orientation dependent hydrogen-bonding potential\textsuperscript{22}, backbone-dependent rotamer probabilities\textsuperscript{23}, a knowledge-based electrostatic energy term\textsuperscript{24}, amino acid probabilities conditioned on φ and ψ space\textsuperscript{25} and reference energies that approximate the unfolded state energy of an amino acid.\textsuperscript{26}

**Plasmid constructs, gene synthesis and mutagenesis**

Kinase dead mutants (K298R) of full length PAK1 and PAK1\textsuperscript{V127E/S144E} were cloned in pQE-80 L (Qiagen) vector. Kinase domain (K298R mutant), residues 250-545 from PAK1 was also cloned in pQE-80L vector with an extra sequence (MRGSHHHHHHHGSDYDIPTTENLYFQC) in N-terminus. The PAK1 kinase\textsuperscript{L470E} mutant was made by overlap extension using PCR.\textsuperscript{27} Kinase domain of PAK5, residues 425-718 was cloned in pGEX-4T-1 vector with an extra sequence (MHHHHHHSSGVDLGTENLYFQSM) in N-terminus.

WT_IS, Hairpin and PAcKer were expressed were made as an MBP fusion with a TEV protease cleavage site and cloned in pQE-80L vector so that 6×His-tag remained at N-terminus after the expression. AlmostHelix was cloned in pGEX-4T-1 vector so that the protein was expressed as GST fusion with a Thrombin cleavage site. CFP-PAcKer was cloned in pQE-80L vector with an N-terminus 6×His-tag.
Expression and protein purification

All expressions were carried in BL21(DE3) pLysS cells. The cells were grown upto OD<sub>600</sub> 0.6-0.8 and then induced with 0.3 mM IPTG and further grown at 25 °C for 6 h for protein expression. The only exception GST-PAK5 kinase was expressed overnight at 20 °C. Cells were then disrupted using sonication and resulting lysates were cleared by two rounds of centrifugation (18,000 × g) of 20 mins each. The supernatants having PAK1 variants and CFP-PAcKer were then purified using a prepacked Ni-NTA column (HisTrap, HP, GE Healthcare) followed by an anion exchange step (Source 15Q beads, GE Healthcare). WT_IS, Hairpin and PAcKer were purified using Ni-NTA column followed by an overnight proteolysis using TEV protease and then again loading the product on Ni-NTA column. The second Ni-NTA affinity step removed the 6×His-tagged MBP to give design variants (all with an extra GS sequence at N-terminus) as a flow through. Flow through was concentrated and then loaded on gel filtration column (Superdex 75, GE Healthcare). All the design variants eluted at similar elution volume (size corresponding to monomer) and appropriate fractions combined and concentrated (Amicon Ultra, Millipore). Protein concentration was estimated using theoretical molar extinction coefficients and absorbance at 280 nm. CFP-PAcKer concentration was estimated using CFP absorbance at 433 nm with an extinction coefficient of 44000 M<sup>-1</sup> cm<sup>-1</sup>.

GST tagged proteins, Almost Helix and PAK5 kinase domain were purified using a prepacked column (GSTrap, HP, GE Healthcare). For AlmostHelix, GST tag was cleaved off using Thrombin protease (Sigma-Aldrich). Post cleavage, the sample was loaded on anion exchange column (HiTrap Q, GE Healthcare). AlmostHelix was collected as a flow-through.
GST-PAK5 kinase was further purified using an anion exchange step (Source 15Q beads, GE Healthcare).

**Circular dichroism (CD)**

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans (250-200 nm) were carried at a typical protein concentration of 20 µM in a 1 mm cuvette. The temperature was maintained at 20°C using a Peltier device. Variable temperature scan was carried between 4-96°C while measuring the ellipticity at 222 nm. All ellipticity data were corrected with a buffer blank and then converted to mean residue ellipticity. Osmolyte TMAO was dissolved in the same buffer and added to the protein. Ellipticity was subtracted from the buffer blank containing same amount of TMAO.

**Isothermal titration calorimetry**

The binding affinity of the designed proteins was measured using VP-ITC isothermal titration calorimeter (MicroCal, GE Healthcare). The designs were used at concentrations of 14-21 µM (AlmostHelix at a concentration of 50 µM) in the cell. PAK1 variants were used as titrants at a 12-15 fold higher concentrations. The proteins were exhaustively dialysed in 20 mM Na-Phosphate buffer, pH 7.4, 25 mM NaCl and 5 mM 2-mercaptoethanol (BME). A total of 29 titrations of 10 µl each were made. The data was analyzed using Origin50 software and fitted using a model for ‘one-site binding’.

**Fluorescence polarization assay**

The binding affinity of AlmostHelix was also measured using fluorescence polarization technique which has been described earlier. S36C mutation was made on AlmostHelix for
conjugation of thiol-reactive fluorescent probe Bodipy(507/545)-iodoacetamide (Molecular Probes). Fluorescence polarization assays were carried out on a Jobin Yvon Horiba Spex FluoroLog-3 instrument (Jobin Yvon Inc) performed in L-format with the excitation wavelength set at 508 nm and emission wavelength set at 545 nm. Bodipy conjugated AlmostHelix (in 50 mM Tris-Cl, pH 7.5, 5 mM BME) at a final concentration of 5 µM and volume 180 µl was titrated with PAK1 kinase domain (in 50 mM Tris-Cl, pH 7.5, 50 mM NaCl, 5 mM BME).

**Conjugation of fluorescent dye to CFP-PAcKer**

Solvent sensitive merocyanine dye was conjugated to CFP-PAcKer to create a biosensor (mero-CFP-PAcKer). Freshly reduced CFP-PAcKer with K→C mutation (at position equivalent to 134 in PDB 1F3M) was buffer exchanged to 50 mM Na-Phosphate (pH 7.5) using PD10 desalting column (Amersham Biosciences). 3-5 equivalents of mercyanine dye dissolved in DMSO (dimethyl sulfoxide) was added to 1 ml aliquot of protein (typical concentration 40-100 µM) in an eppendorf tube and immediately mixed by inverting the tube several times. The tube was then covered with aluminium foil and left under gentle mixing for 2 h at room temperature. The conjugation reaction was then terminated using excess of BME (5 µl stock in 1ml reaction mix). Unreacted dye was separated from the conjugated protein using PD10 column equilibrated and then eluted with 20 mM Na-Phosphate (pH 7.4), 25 mM NaCl, 5 mM BME. Dye-labeled protein concentration was estimated using CFP absorbance at 433 nm (ε = 44000 M⁻¹ cm⁻¹). Dye concentration was estimated using dye absorbance at 593 nm (ε = 120000 M⁻¹ cm⁻¹). The conjugation efficiency (Dye concentration / CFP concentration) observed was ~100%.
Fluorimetric measurement of mero-CFP-PAcKer interaction with PAK1

Dye-labeled CFP-PAcKer (mero-CFP-PAcKer) in an assay buffer (20 mM Na-Phosphate, pH 7.4, 25 mM NaCl, 5 mM BME) was titrated with PAK1 variants so that final concentration of mero-CFP-PAcKer was 200 nM and of PAK1 variants ranged between 0-10 µM. The reaction mix for each titration was made in separate eppendorf tube and incubated for 10 mins at room temperature. Readings were taken on Jobin Yvon Horiba Spex FluoroLog-3 instrument (Jobin Yvon Inc) where the sample in 3 mm cuvette was excited with 433 nm and 593 nm and emission signal captured at 474 nm (CFP signal) and 620 nm (merocyanine dye signal) respectively. The readings were taken from lower titrant concentration to higher without an intermediate cuvette washing step. Typically data accumulation took overall 30 mins from the start of experiment (0 titrant reading) to end of experiment (10 µM titrant reading). A normalized fluorescence ratio \(((\text{Dye/CFP})_{\text{Titrant}} / (\text{Dye/CFP})_{\text{Buffer}})\) was calculated and plotted against titrant concentration. The data was fit using SigmaPlot using Michaelis-Menten equation for ‘one-site binding’.

For specificity experiment GST-PAK5 kinase domain was titrated to a final concentration of 1 µM. For a proper comparison, PAK1 kinase domain at 1 µM was supplemented with 1 µM of GST protein.
Supplementary figure S2.1. AlmostHelix gains structure in the presence of TMAO. (a) Far-UV CD scan of AlmostHelix shows increase in ellipticity at 222 nm as a function of TMAO concentration. (b) Thermal denaturation of AlmostHelix at 222 nm shows steeper transition and increase in Tm at 2M TMAO confirming gain in conformational stability.
Supplementary figure S2.2. Isothermal titration calorimetry assay to determine the binding affinity of CFP-PAcKer against ‘open’ form of PAK1. 262 µM PAK1 V127E/S144E mutant (model for full length ‘open’ form) was titrated against 21 µM CFP-PAcKer. The binding affinity was weaker than PAK1 kinase domain (model for active conformation).
Supplementary figure S2.3. A destructive mutation on either side of proposed CFP-PAcKer and kinase domain interaction resulted in complete loss of signal in ITC experiments. (a) 208 µM of PAK1 kinase domain was titrated in 18 µM of CFP-PAcKer<sup>V127E</sup>. (b) 208 µM of PAK1 kinase<sup>L470E</sup> mutant was titrated in 18 µM of CFP-PAcKer.
Supplementary figure S2.4. Mutant biosensor (CFP-PAcKer\textsuperscript{V127E} based) against PAK1 kinase domain. The normalized fluorescence ratio did not change with titration.

Supplementary figure S2.5. ITC assay to detect interaction between CFP-PAcKer and GST-PAK5 kinase domain. 223 µM GST-PAK5 kinase was titrated into 20 µM CFP-PAcKer. Very weak signal (endothermic) was observed.
Supplementary figure S2.6. GST-PAK5 kinase shows poor interaction with the CFP-PAcKer derived biosensor but shows a different saturation. Since PAK5 kinase was GST tagged, equimolar amount of GST was added in PAK1 kinase titration for a better comparison.
Supplementary figure S2.7. ITC assay to show CFP-IS and PAK1 kinase domain interaction. 208 µM of PAK1 kinase domain was titrated into 18 µM of CFP-IS. The binding affinity was 1.65 µM with a molar ratio of ~1.
References


CHAPTER 3

COMPUTATIONAL DESIGN OF A PAK1 BINDING PROTEIN

Abstract

We describe a computational protocol, called DDMI, for redesigning scaffold proteins to bind to a specified region on a target protein. The DDMI protocol is implemented within the Rosetta molecular modeling program and uses rigid-body docking, sequence design, and gradient-based minimization of backbone and side chain torsion angles to design low energy interfaces between the ‘scaffold’ and ‘target’ protein. Iterative rounds of sequence design and conformational optimization were needed to produce models that have calculated binding energies that are similar to binding energies calculated for native complexes. We also show that additional conformation sampling with molecular dynamics can be iterated with sequence design to further lower the computed energy of the designed complexes. To experimentally test the DDMI protocol we redesigned the human hyperplastic discs protein to bind to the kinase domain of p21-activated kinase 1 (PAK1). Six designs were experimentally characterized. Two of the designs aggregated and were not characterized further. Of the remaining four designs, three bound to the PAK1 with affinities tighter than 350 µM. The tightest binding design, named Spider Roll, bound with an affinity of 100 µM. NMR–based structure prediction of Spider Roll based on backbone and $^{13}$Cβ chemical shifts using the program CS-ROSETTA indicated that the architecture of human hyperplastic discs
protein is preserved. Mutagenesis studies confirmed that Spider Roll binds the target patch on PAK1. Additionally, Spider Roll binds to full length PAK1 in its activated state, but does not bind PAK1 when it forms an auto-inhibited conformation that blocks the Spider Roll target site. Subsequent NMR characterization of the binding of Spider Roll to PAK1 revealed a comparably small binding ‘on-rate’ constant (<< $10^5$ M$^{-1}$ s$^{-1}$). The ability to rationally design the site of novel protein-protein interactions is an important step towards creating new proteins that are useful as therapeutics or molecular probes.
Introduction

Protein-protein interactions (PPI) are indispensable for life and irregularities in PPI are implicated in many pathological conditions. The rational design of PPIs is a rigorous test of our understanding of molecular recognition and accurate design strategies should allow for the creation of novel protein therapeutics, diagnostics and research tools. Recently there has been considerable success in the computational redesign of protein binding affinities and specificities.\textsuperscript{1, 2, 3} In these studies, rotamer and sequence optimization protocols have been used to identify amino acids that form good packing interactions, electrostatic interactions and hydrogen bonds at target interfaces. In general, these simulations begin with a high-resolution crystal structure of the target interaction. Considerably more difficult is the design of protein interactions for which there is no starting structure. There have been impressive results in the design of new coiled-coils, but these studies rely on known patterns of recognition between coiled-coils.\textsuperscript{4, 5, 6} The rational design of novel interfaces between arbitrarily chosen proteins remains largely an unsolved problem.

Recent successes in directed evolution of PPI indicate that even fairly rigid protein scaffolds can be remodeled to bind new target proteins. Ribosome display has been used to design ankyrin repeat proteins that bind with high affinity to maltose binding protein and aminoglycoside phosphotransferase.\textsuperscript{7, 8} Crystal structures of the complexes show only small changes in the conformations of each protein when they dock together. These results suggest
a minimal protocol for computer-based interface design: dock the scaffold on to the target protein and then redesign the amino acids on the surface of the scaffold to form favorable interactions with the target. There are many ways that two proteins can be brought together and some orientations are likely to be more designable than others. The challenge is that before redesigning the surface of the scaffold in the presence of the target, it is difficult to determine which docked orientation will provide the lowest energy interactions. Huang and Mayo used a reduced representation of amino acid side chains and a fast Fourier-transform based docking algorithm to find orientations and positions that maximize potential interactions with the target without bringing the proteins too close together.\textsuperscript{9} They used this strategy to redesign the $\beta_1$ domain of streptococcal protein G to form a novel heterodimer with a binding affinity of $\sim300 \, \mu\text{M}$.\textsuperscript{10} In their study, the sequences of both sides of the protein interface were optimized and only one docked conformation was explicitly evaluated with protein design simulations.

We have developed a strategy for interface design, called DDMI for dock, design and minimize interface, which is based on the premise that it is advantageous to explicitly consider many alternate docked orientations and positions. This protocol builds on Rosetta’s existing fixed-backbone design subroutine\textsuperscript{11}, that, when given a docked conformation for the scaffold and target backbones, searches through side chain sequence- and conformation-spaces to produce a low-energy sequence for the docked conformation. Since we cannot know before design begins, which docked conformations will lead to good sequences, we must sample many docked conformations. We therefore precede the design step with a stochastic, low resolution, rigid-body docking of the two proteins.\textsuperscript{12} Only a single docked conformation from the docking stage is passed into the subsequent design phase; however it
is rare that two docking trajectories produce the same docked conformation. This initial docking phase effectively ensures that the design phase is seeded with a unique docked conformation. After docking completes, DDMI iterates between rounds of design and gradient-based minimization to settle into a low-energy sequence for the scaffold protein. Independent DDMI trajectories settle into dissimilar regions of conformation space, so we typically simulate tens- to hundreds-of-thousands design trajectories. Here we use the DDMI protocol to design an interaction between the hyperplastic discs protein (HYP) and the kinase domain of p21-activated kinase 1 (PAK1).

Results

The model system

We chose kinase domain of p-21 activated kinase 1 (PAK1) as our 'target' protein. The full length PAK1 (PAK1-fl) is a multi-domain protein that can switch between an inactive and active conformation. In the inactivated state, the auto-inhibitory domain of PAK1-fl binds with the kinase domain of PAK1. In the activated state, the auto-inhibitory domain is unfolded by accessory factors and no longer interacts with the kinase domain. For de novo interface design we targeted a region of the PAK1 that is exposed when the auto-inhibitory domain releases. This is an attractive binding site because it is a known region of protein-protein interaction and binders that target this region will be sensitive to the activation state of PAK1, potentially providing a tool for sensing or controlling PAK1 activity. The 'target' patch is a hydrophobic cleft in the C-terminal domain of the kinase domain between two α-helices (αEF and αG). The auto-inhibitory domain of PAK1-fl forms a small helical bundle that inserts in the cleft. As a design scaffold for targeting the PAK1 kinase domain, we used
a small helical bundle protein, the Hyperplastic discs protein (HYP, PDB ID: 1I2T),\(^{14}\) that is similar in size to the auto-inhibitory domain and can fit in the target cleft.

In a preliminary set of HYP designs we found that it was prone to aggregation when redesigned to bind PAK1. Because we are targeting a hydrophobic cleft on PAK1, the redesigned scaffolds typically have additional hydrophobic groups on their surface. To increase the baseline solubility of HYP we selected positions away from the target interface for mutagenesis to polar residues. The mutations, G26E, L37E and L38N are presented in some of the designs discussed here. Additionally, we introduced the mutation A15C to HYP to allow for conjugation of the fluorophore Bodipy for measuring changes in fluorescence polarization upon binding to PAK1. Circular dichroism spectra indicate that these mutations do not perturb the helical structure of HYP (Supplementary figure S3.1).

**Interface design protocol, DDMI**

To redesign HYP to bind the PAK1 we developed an interface design protocol, called DDMI, within the Rosetta molecular modeling suite\(^ {15};\)^\(^ {16}\) (Figure 3.1). The first stage of the protocol was rigid-body docking with a Monte Carlo optimization protocol and a low resolution model of the amino acid side chains.\(^ {12}\) The goal of this stage was to find a plausible docked-conformation for the two proteins so that they can be designed to bind in this conformation. We filter at this stage any docked conformations that show backbone collisions which cannot be removed in the design phase. To bias sampling of the scaffold’s conformation to bind against the target cleft, constraints were added to the energy function to reward the burial of the residues in the cleft. These constraints were then removed after the initial stage of docking. Figure 3.2 (a) shows the representative sampling of structures created during ‘dock’ stage of the protocol.
Figure 3.1. **DDMI protocol for protein interface design.** Each trajectory starts with rigid-body docking using a low resolution score function with additional constraints to direct predetermined residues to the interface. A docked conformation with a high score (binding energy > 18 REU with a low resolution energy function) was rejected (Filter I). When a trajectory passes the filter, it goes through 8 rounds of sequence optimization and backbone minimization during which LJ-repulsive are ramped upwards and coordinate constraints are ramped downwards for 6 rounds. During this process the Lennard-Jones repulsive term of the energy function is ramped from 0.85 to 1, whereas the coordinate constraint weight goes from 1 to 0. Designs are filtered before being output based on a binding energy density (<-0.01 REU/Å²) and a maximum number of 4 unsatisfied polar groups at the interface (Filter II).
Figure 3.2. Conformational sampling in (a) ‘dock’ stage and (b) energy convergence during the ‘design/minimize’ stage. In (a) the region in ‘magenta’ depicts the ‘hotspot’ region on the target protein. In (b) each line represents an independent trajectory. A drop in energy is observed after cycles 11 and 12 during which the coordinate constraints are ramped to the final value and are sufficiently weak to let the proteins relieve the strain induced while holding the two partners next to each other.

After docking, iterative rounds of sequence design and structure optimization with Rosetta’s all-atom energy function\textsuperscript{16} were used to find low energy sequence structure pairs.
Sequence design was performed with simulated annealing and a rotamer-based representation of the amino acid side chains. Structure optimization was performed using gradient-based minimization of rigid-body orientation as well as backbone and side chain torsion angles. In the early rounds of design and minimization, DDMI weakened the repulsive component of the Lennard-Jones potential and added coordinate constraints for the backbone Cα atoms to prevent the fledgling interface from ‘exploding’: the typical binding energy for the interfaces that resulted from the first few iterations was positive. Without the coordinate constraints, the minimizer displaced the scaffold from the target, preventing the design of any interface. In each iteration, DDMI decreased the weight on the coordinate constraints and increased the weight on the Lennard-Jones repulsive term. We found that after eight rounds of sequence design and structure minimization most trajectories converged on a local minimum (Figure 3.2 b).

Global sampling of conformational space (within the target constraints) was achieved by performing independent trajectories that start from uniquely docked complexes. We performed >1 million independent DDMI trajectories with HYP and PAK1. To assess the quality of our designs, we compared our models against 43 naturally occurring protein-protein interfaces with high-resolution crystal structures (2.3 Å or less, Supplementary table S3.1). We observed three key differences between native interfaces and designed interfaces. In the native interfaces, the minimized Rosetta energies correlated strongly with the interface sizes (Supplementary figure S3.2). The average binding energy density, defined as the binding energy (in Rosetta Energy Units, REU) per buried surface area (Å²) was -0.013 REU/Å² for the native interfaces (Table 3.1). Binding energy density was a potent discriminator of designed and natural interfaces; most DDMI trajectories resulted in
interfaces with poor binding energy densities. We also observed that native structures had on average 4 unsatisfied hydrogen bonds at their interfaces; whereas DDMI models often contained many more. Finally, we observed that naturally occurring interfaces were packed more tightly than those produced by DDMI as measured by the SASApack score. The SASApack score in Rosetta is derived from examining the difference in the molecular surface areas accessible to a 0.5 Å radius probe and accessible to a 1.4 Å radius probe.\textsuperscript{17} This difference indicates surface area on the protein that is not in contact with either water or other protein atoms, and hence reflects the presence of voids that are too small to be filled with water. The score is normalized by the average surface-area difference observed in a large set of crystal structures. A negative SASApack score indicates better packing than crystal structures, a positive SASApack score indicates worse packing. The average SASApack score for the set of native interfaces was -1.39 ± 1.29, whereas many of the designs had positive SASApack scores.

| Table 3.1. Benchmark scores from native PDBs and the cutoffs used as Filter II during simulations. |
|-------------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Buried surface area (Å$^2$) | Polar surface area (Å$^2$) | Binding energy density (REU/Å$^2$) | Unsatisfied polar groups | SASApack score |
| Native complexes | 2060 ± 675 | 637 ± 249 | -0.013 ± 0.003 | 4.26 ± 2.44 | -1.39 ± 1.29 |
| Filter II | - | - | -0.01 | 4 | - |

These observations motivated the set of filters we incorporated into the DDMI protocol. DDMI discarded designs if their binding-energy density was higher than -0.01 REU/Å$^2$ and if they buried 4 or more polar groups which lacked hydrogen bonding partners.
Final selection of the design models was made based on the SASApack score of less than 2.0 and a minimum buried surface area of 700 Å². Satisfying all four criteria required considerable sampling; less than 1% of the DDMI trajectories passed these filters. The interfaces that were left were of moderate size (900 Å² – 1700 Å²).

Four designs with favorable values for all of the evaluation metrics were selected for experimental validation. Each of them had the c-terminal helix of the scaffold interacting with the ‘hotspot’ cleft in PAK1. The main interacting residues on PAK1 were L470, L473 and Y474. In some cases V436, R438 and R471 were also contributing to the interaction. On the scaffold side, the mutations were mainly concentrated on the helix IV (Figure 3.3). Residues involved in interaction were mostly hydrophobic consisting of one or two aromatic amino acids forming the centre of interactions. Design model 1212 had the greatest number of polar residues at the interface, and design model s032 had the fewest polar amino acids at the interface (Table 3.2).

Figure 3.3. Multiple sequence alignment of the design models selected for experimental validation. The interface residues in each case are highlighted.
Table 3.2. Computational scores of the selected design models and comparison with a native AID-kinase interaction in PAK1 full length.

<table>
<thead>
<tr>
<th>Design</th>
<th># of mutations on 'scaffold'</th>
<th># of Dock-design trajectories</th>
<th>Buried surface area (Å²)</th>
<th>Polar buried surface area (Å²)</th>
<th>Binding energy density (REU/Å²)</th>
<th># of unsatisfied polar groups</th>
<th>SASApack score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0233</td>
<td>9</td>
<td>3000</td>
<td>1507</td>
<td>286</td>
<td>-0.0118</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>Spider Roll</td>
<td>12</td>
<td>10000</td>
<td>966</td>
<td>174</td>
<td>-0.0152</td>
<td>0</td>
<td>-0.65</td>
</tr>
<tr>
<td>1212</td>
<td>10</td>
<td>20000</td>
<td>1500</td>
<td>383</td>
<td>-0.0134</td>
<td>0</td>
<td>2.31</td>
</tr>
<tr>
<td>3533</td>
<td>10</td>
<td>1681</td>
<td>323</td>
<td>-0.0129</td>
<td>3</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>s032a</td>
<td>15</td>
<td>320</td>
<td>983</td>
<td>124</td>
<td>-0.0167</td>
<td>0</td>
<td>-3.26</td>
</tr>
<tr>
<td>s037a</td>
<td>21</td>
<td>370</td>
<td>1207</td>
<td>288</td>
<td>-0.0173</td>
<td>0</td>
<td>0.72</td>
</tr>
<tr>
<td>Trunc-AID</td>
<td>-</td>
<td>-</td>
<td>1619</td>
<td>357</td>
<td>-0.0157</td>
<td>2</td>
<td>-0.027</td>
</tr>
</tbody>
</table>

a: The models were generated using discrete molecular dynamics on Spider Roll for different backbone conformations and the binding interface was designed using DDMI.

Discrete molecular dynamics (DMD) to sample more backbone conformations

To study the effect of additional backbone flexibility on the design process, we used discrete molecular dynamics\textsuperscript{18; 19; 20} to sample conformational space near one of the designed complexes (Spider Roll). From the initial design model, we first performed short DMD simulations and picked ten structure snapshots from the simulation trajectory. Next, we applied sequence design and minimization (no docking) from the DDMI protocol to the snapshots to generate ten new designs, from which the one with the best Rosetta energies was selected for the next round of DMD simulation. The iterations were repeated 100 times at decreasing simulation temperature to identify low energy designs. Finally we examined all the designs that were generated from the iteration protocol, from which we selected two designs, named s032 and s037, for experimental verification. The backbone RMSD between s032 and Spider Roll when the target protein was superimposed was 5.36 Å while between s037 and Spider Roll the deviation was 4.53 Å. The sequences of the two new designs differ
significantly from the initial Spider Roll design. Out of the 21 residue sites being designed, the numbers of mutated amino acids are 12 and 18 for s032 and s037, respectively.

**Binding measurements and 'hotspot' mapping**

The six designs were expressed as MBP fusions. All of them expressed in the soluble fraction of *E. coli* lysate, but 3533 and s032 aggregated when MBP was removed with TEV protease. No further studies were performed with 3533 and s032. The circular dichroism (CD) spectra of the designs indicate that the proteins are helical and all of the designs exhibited cooperative thermal melts as monitored by the CD signal at 222 nm (Supplementary figure S3.1). The four soluble designs were labeled with the fluorescence probe Bodipy, and fluorescence polarization was used to monitor binding to PAK1 (Table 3.3). Spider Roll showed the best binding affinity (Kₐ = 100 µM) while design model s037 which was derived from Spider Roll using DMD bound with a dissociation constant of 160 µM. Model 1212 bound with an affinity of 330 µM and model 0233 failed to show any conclusive binding with PAK1.

<table>
<thead>
<tr>
<th>Design</th>
<th>Gel Filtration</th>
<th>Far-UV CD Scan</th>
<th>Thermal denaturation</th>
<th>FP Assay: Kₐ (µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0233</td>
<td>Monomer</td>
<td>Folded</td>
<td>Cooperative</td>
<td>Poor fit</td>
</tr>
<tr>
<td><strong>Spider Roll</strong></td>
<td>Monomer</td>
<td>Folded</td>
<td>Cooperative</td>
<td>100</td>
</tr>
<tr>
<td>1212</td>
<td>Monomer</td>
<td>N/D</td>
<td>Poor fit</td>
<td>330</td>
</tr>
<tr>
<td>3533</td>
<td>Aggregate</td>
<td>N/D</td>
<td>Poor fit</td>
<td>160</td>
</tr>
<tr>
<td>s032</td>
<td>Aggregate</td>
<td>N/D</td>
<td>Poor fit</td>
<td>160</td>
</tr>
<tr>
<td>s037</td>
<td>Monomer</td>
<td>Folded</td>
<td>Cooperative</td>
<td>160</td>
</tr>
</tbody>
</table>
To probe if Spider Roll was interacting with PAK1 as designed (Figure 3.4), we mutated residues on both sides of the interface that were predicted to contribute to binding.

**Figure 3.4. Spider Roll – PAK1 model showing the interface residues.** On the PAK1 side (yellow) L470, R471, L473 and Y474 are the main residues involved in interaction. On the Spider Roll side the main contributions come from Y52, A55, G56, I58 and F59. G61 is involved in hydrogen bonding with R438.

Consistent with the design model, the PAK1 mutations L473A and L470E each destabilized binding by over 1.4 kcal/mol (Figure 3.5 a). R471A and Y474A also destabilized binding by 0.35 and 0.85 kcal/mol respectively. R438, predicted to form a hydrogen bond with backbone carbonyls at the C-terminus of Spider Roll, did not have any effect on binding affinity when mutated to alanine. Mutations to residues on Spider Roll that were designed to form interactions with PAK1 also weakened binding (Figure 3.5 b). The point mutations F59A and G56E weakened binding by 0.55 kcal/mol. Y52A and A55E mutations had only subtle effect with a binding energy cost of approximately 0.4 kcal/mol. A more dramatic decrease in binding was observed by combining F59A with G56E (1.4 kcal/mol). I58A mutation, which was predicted to contact PAK1 at the edge of the interface, had very
minimal effect on binding. The design scaffold (with solubilizing mutations G26E, L37E and L38N) had basal level of affinity for PAK1. Taken together the mutational data indicates that Spider Roll interacts with the target patch on PAK1, and that the helix IV of Spider Roll is involved in binding.

Figure 3.5. Mutagenesis studies of Spider Roll – PAK1 interface. (a) Effects of Spider Roll mutations on Spider Roll – PAK1 interaction in FP assays. HYP mutant (scaffold with A15C, G26E, L37E and L38N) did not show any binding. (b) Effects of PAK1 mutations on Spider Roll – PAK1 interaction. L470E and L473A mutations showed maximum disruption (more than 10 fold weakening in binding). R471A and Y474A had subtle effects while R438A did not show any effect on interaction.
**NMR-based structure prediction of Spider Roll**

In order to structurally characterize Spider Roll as well as its binding to PAK1, we nominated Spider Roll mutant I58A, which expressed with a more than two-fold increases yield when compared with Spider Roll (see Methods), as a community outreach target of the Protein Structure Initiative (PSI) 2 and collaborated with the Northeast Structural Genomics Consortium (NESG: [http://www.nesg.org](http://www.nesg.org)) to obtain sequential polypeptide backbone and $^{13}\text{C}\beta$ chemical shift assignments. The 2D-$^{15}\text{N},^{1}\text{H}$ HSQC spectrum recorded for Spider Roll I58A shows favorable chemical shift dispersion and indicates that the protein is folded in solution (Supplementary figure S3.3). Assignment completeness of detectable peaks in the 2D-$^{15}\text{N},^{1}\text{H}$ HSQC spectrum was 91% (48/53). Polypeptide backbone and $^{13}\text{C}\beta$ chemical shift assignments (Supplementary table S3.2) were obtained for 50 residues and a total of 81% of the shifts assignable with the selected set (see Methods) of multidimensional NMR experiments (i.e. excluding the N-terminal $^{15}\text{NH}_3^+$, the three prolyl $^{15}\text{N}$ and the $^{13}\text{C}'$ shifts of residues preceding prolyl residues). The chemical shifts are in agreement with the location of $\alpha$-helices in the X-ray crystal structure of the design scaffold protein (HYP), except for the last ~6 residues of the helix IV (Supplementary figure S3.4, Supplementary table S3.3). The shifts were then used to predict the structure of Spider Roll I58A with the program CS-ROSETTA$^{22}$ (Supplementary figure S3.5). The CS-ROSETTTA structure is very similar structure of design template protein HYP (backbone RMSD = 0.7 Å for helical residues; residues 1010-1022, 1026-1036, 1041-1049 and 1051-1065 for 1I2T and residues 2-14, 18-28, 33-41 and 43-57 for the current structure) which indicates that the re-design of HYP did not significantly affect the fold of the protein.
The chemical shift indices suggest that the last ~6 residues of the C-terminal helix of Spider Roll are frayed in solution. To further investigate the conformation of the helix IV, we derived amide proton – amide proton upper distance limit constraints from 3D $^{15}\text{N}$-resolved [$^{1}\text{H}, ^{1}\text{H}$]-NOESY. The longer distances derived for the C-terminal segment of helix IV reflect weaker NOEs (the sequential NOEs between the last 6 residues either overlap or disappear), which are consistent with fraying of this segment. All of our designed interfaces include a fully intact C-terminal helix that makes close contact with PAK1. Fraying of the helix IV in the unbound state is not inconsistent with a fully folded helix in the bound state, but indicates that there will be an additional entropic penalty associated with binding. Hence, future improvement of the design of Spider Roll may focus on stabilizing the C-terminal segment of helix IV.

**NMR characterization of Spider Roll-PAK1 binding**

The 2D [$^{15}\text{N}, ^{1}\text{H}$] HSQC spectrum$^{21}$ was recorded for Spider Roll I58A was monitored as a function of PAK1 concentration. The spectra were recorded at three different molar ratios of Spider Roll I58A / PAK1: 231 µM / none, 210 µM / 120 µM, 148 µM / 487 µM). Unexpectedly, addition of PAK1 did not induce any perturbation of Spider Roll chemical shifts or introduce large broadening of Spider Roll resonances, but lead to a dramatic decrease of Spider Roll signal intensities. Specifically, at a PAK1 concentration of 487 µM, Spider Roll peak intensities were reduced to <5% of their starting values (Figure 3.6). Furthermore, no ‘new’ signals emerged during titration, which could still be attributed to Spider Roll bound to PAK1. This is likely due to the fact that the signals of the 43 kDa complex present at a concentration of only ~200 µM are too weak to be detected. Assuming for simplicity that both $^{1}\text{H}$ and $^{15}\text{N}$ line-widths scale linearly with molecular weight$^{21}$, the
S/N ratios for Spider Roll in complex with PAK1 are expected to be about 40-fold lower than for free Spider Roll.\textsuperscript{23} Since the S/N ratios observed for free Spider Roll are in the range of ~20 in the 2D-[\textsuperscript{15}N, \textsuperscript{1}H] HSQC spectra, detection of the signal of bound Spider Roll is indeed not be expected.

Furthermore, Spider Roll may bind to PAK1 in different and slowly exchanging conformations, a phenomenon which would broaden resonance lines, thereby further impeding signal detection for bound Spider Roll. A third scenario which would manifest itself by lack of signals for the bound protein would be the formation of aggregates formed by the complex.

To check if the Spider Roll-PAK1 complex aggregates non-specifically, we performed size exclusion chromatography with the NMR sample and also ran an SDS page gel for an aliquot of the NMR sample. The SDS gel confirmed that the same amount of Spider Roll remained in solution as was initially added to the NMR sample, and the size exclusion chromatography indicated that the sample was not aggregating: the only two peaks in the chromatogram corresponded to monomeric PAK1 and Spider Roll, which is typically observed for proteins with micromolar binding affinities. Hence, these experiments suggest that the absence of NMR peaks from the bound state is indeed likely due to the slower overall rotational tumbling of the complex.

Our NMR data are consistent with the lifetime of the Spider Roll-PAK1 complex being much longer than the time required for signal detection (‘NMR chemical shift time scale’, around 0.1 s), that is, we obtain as an upper bound for the ‘off-rate’ $k_{off} << 10 \text{ s}^{-1}$. With the dissociation constant $K_D = k_{off}/k_{on} = 10^{-4} \text{ M}$, we then obtain for the ‘on-rate’ $k_{on} << 10^5 \text{ M}^{-1} \text{ s}^{-1}$. The on-rate constants for protein-protein binding can vary dramatically ($1\times10^3$
$M^{-1} s^{-1}$ to $1 \times 10^9 M^{-1} s^{-1}$) but are often near $1 \times 10^6 M^{-1} s^{-1}$. The comparably small on-rate constant for Spider Roll could reflect the absence of an effective docking funnel or a conformational change that accompanies binding, such as folding of the end of the helix IV in Spider Roll.

![Normalized Spider Roll I58A HSQC peak volumes as a function of PAK1 concentration](image)

**Figure 3.6. Normalized Spider Roll I58A HSQC peak volumes as a function of PAK1 concentration.** Titration 1: 231 µM Spider Roll I58A with no PAK1; Titration 2: 210 µM Spider Roll I58A with 120 µM PAK1 (green bar) and 213 µM Spider Roll I58A with 124 µM PAK1 mutant (red bar); Titration 3: 148 µM Spider Roll I58A with 487 µM PAK1 (green bar) and 150 µM Spider Roll I58A with 490 µM PAK1 mutant (red bar). Positive and negative error bars are one standard deviation.

To validate that the reductions of Spider Roll peak intensities in 2D-[\(^{15}\)N, \(^1\)H] HSQC are due to binding to the kinase domain, the NMR titration was repeated for PAK1 mutant L470E, which shows reduced affinity for Spider Roll when monitored using fluorescence polarization experiments. Consistently, the changes in peak volume were much smaller than those observed with wild type PAK1: at a PAK1 L470E concentration of 470 µM, the Spider
Roll I58A peak volumes were still 40% of their original size (in contrast to <5% for WT PAK1) (Figure 3.6, Supplementary figure S3.6).

**Does Spider Roll adopt multiple docked positions when binding PAK1?**

Both the NMR data and mutational data indicate that Spider Roll binds the target cleft on PAK1, but they do not rule out the possibility that it can adopt alternative docked orientations when bound to PAK1. To further examine this possibility we used Rosetta to perform protein-protein docking simulations with Spider Roll and PAK1. In these simulations, Spider Roll was constrained to be near the target binding site, but was allowed to adopt alternative orientations relative to PAK1. Many independent trajectories were used to probe the energy landscape and the energies of the various models were plotted versus RMSD to the target conformation. We identified two clusters of low energy structures (Figure 3.7 a). The lowest energy cluster was centered on the design model, but the second cluster packed helix IV in a direction that was orthogonal to the design model (Figure 3.7 b, c). The mutational data does not strongly distinguish between the two alternatives. The mutations that have the strongest effect on binding energy are buried in both sets of models (Table 3.4) as calculated by the NACCESS program.26

Table 3.4. Buried solvent accessible surface area (SASA) of the interface residues in Spider Roll-kinase complex in two alternate docked positions (Figure 3.7). NACCESS program26 was used to calculate the absolute SASA of each interface residue in the complex and then subtracted from SASA of each residue in the independent chains to give buried SASA.

<table>
<thead>
<tr>
<th>Interface residue</th>
<th>R438</th>
<th>L470</th>
<th>R471</th>
<th>L473</th>
<th>Y474</th>
<th>Y52</th>
<th>A55</th>
<th>G56</th>
<th>I58</th>
<th>F59</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low RMSD (Å²)</strong></td>
<td>81</td>
<td>104</td>
<td>43</td>
<td>50</td>
<td>98</td>
<td>102</td>
<td>52</td>
<td>22</td>
<td>53</td>
<td>135</td>
</tr>
<tr>
<td><strong>High RMSD (Å²)</strong></td>
<td>110</td>
<td>135</td>
<td>49</td>
<td>61</td>
<td>67</td>
<td>125</td>
<td>35</td>
<td>16</td>
<td>27</td>
<td>113</td>
</tr>
</tbody>
</table>
Figure 3.7. Presence of two competing states in Spider Roll – PAK1 model. (a) Energetic ‘dock’ funnel shows the presence of two competing states, designated by two stems of the funnel. (b) Orientation of interface residues in low RMSD confirmation (PAK1 in yellow and Spider Roll in cyan). (c) Orientation of same residues in high RMSD state (PAK1 in yellow and Spider Roll in magenta). Ineffectiveness of I58 and R438 mutations (in Figure 3.5) is consistent with the State 2 where there is no involvement of these residues in interaction.

Spider Roll binds preferentially to the activated form of full-length PAK1

In its inactive form, full-length PAK1(PAK-fl) forms a closed conformation in which an autoinhibitory domain binds to the same cleft in the kinase domain that we have targeted with Spider Roll. PAK1 can be opened by introducing mutations in the autoinhibitory elements (V127E, S144E) that weaken affinity for the kinase domain. Using fluorescence polarization we measured the affinity of Spider Roll for WT PAK-fl and PAK1-fl with the
mutations V127E and S144E. Spider Roll showed no binding with WT inactive PAK1, but bound the activating mutant with an affinity of 200 μM (Figure 3.8).

![Graph showing binding titrations with PAK1 V127E/S144E mutant and PAK1-fl (inactive form).](image)

**Figure 3.8.** Spider Roll distinguishes between ‘closed’ and ‘open’ full-length PAK1 (PAK1-fl). Spider Roll binding titrations with PAK1 V127E/S144E mutant (PAK1-fl mutant, model for full length ‘open’ form) and PAK1-fl (inactive ‘closed’ form).

**Discussions**

The computational design of novel protein-protein interfaces with high specificity and affinity remains an unsolved problem. Here we have explored the effectiveness of an interface design protocol that includes rigid-body docking, rotamer-based sequence design and minimization of side chain and backbone torsion angles. With this protocol we were able to design small interfaces between HYP and PAK1 (∆SASA ~1000 Å²) that had calculated binding energies comparable to naturally occurring interfaces of the same size. Design models with interfaces larger than 1500 Å² either exhibited inferior packing, buried too many
unsatisfied hydrogen bonding groups or had significantly worse binding energy densities than native complexes. These results suggest that more extensive backbone sampling and/or screening of multiple scaffolds will be needed to build larger interfaces with tight packing and good hydrogen bonding.

Three of the six experimentally characterized designs bound PAK1. Spider Roll, which was the tightest binder, bound PAK1 with an affinity of 100 µM. For comparison, the truncated autoinhibitory domain (trunc-AID, residues 83-137) from PAK1 binds to the same target cleft with an affinity of 4 µM (unpublished data). Even though the AID is smaller (55 residues) than the HYP scaffold (61 residues), the interface between trunc-AID and PAK1 is larger than the Spider Roll interface, 1620 Å² versus 970 Å². The calculated binding energy for the trunc-AID-PAK1 complex (after minimization with Rosetta) is -25 REU (-0.016 REU / Å²) compared to -15 REU (-0.015 REU / Å²) for Spider Roll. The more favorable energies achieved by the AID domain (experimental and computational) may reflect the evolution of this domain to have a backbone conformation suitable for binding the PAK1 kinase domain, while the backbone conformation of HYP has not been optimized for binding PAK1.

Mutation studies and NMR studies do not rule out the possibility that Spider Roll samples multiple docked positions when bound to PAK1. This may partially reflect the hydrophobic nature of the designed interface; the buried SASA is 82% nonpolar. For native protein-protein interfaces (Supplementary table S1) the average is 69% nonpolar (or 31% polar, Table 1). Polar interactions can increase specificity because of the strong distance and orientation dependence of hydrogen bond energies. The interfaces designed in this study were predisposed to be hydrophobic because of the hydrophobic nature of the target cleft; however, several polar residues on PAK1 surround the cleft. The AID domain forms
hydrogen bonds with N468 and R471 on PAK1, while these residues are left solvent exposed in the Spider Roll design. Designing multiple hydrogen bonds across a protein interface is challenging because optimizing one bond with rigid-body or backbone perturbations is likely to disrupt the surrounding bonds.

Spider Roll binds to the open state of full length PAK1 but does not bind to the closed inactive form of the protein. Affinity reagents that distinguish the activation state of target molecules are useful starting points for creating biosensors. Typically the proteins are modified with fluorescent groups (chemically or genetically encoded) that change their fluorescent properties when the affinity reagent binds the target. Our results suggest that computational design may be one route for identifying novel affinity reagents. However, to have useful biosensors it will be necessary to have binding affinities tighter than 100 µM. A potential strategy for achieving this goal is to combine computational design with affinity maturation via techniques such as phage or yeast display. This strategy was recently used to improve the rate of catalysis by a computationally designed enzyme.27

Materials and Methods
Computational methods
Rosetta

Our interface design protocol, DDMI, is described in the Results section and Figure 3.1. It was implemented within the Rosetta molecular modeling program.15; 16 A single trajectory consists of a ‘dock’ stage followed by several iterations of ‘design’ and ‘minimization’ stages. The ‘dock’ stage involves a low resolution rigid body Monte Carlo search,12 where the side chain of each residue was represented as one bead placed at centroid position of the side chain. We applied constraints so that the ‘scaffold’ remains close to and packs one or
more of L470, L473 and Y474 residues on the ‘target’. A trajectory satisfying the constraints was designed with 8 iterations of sequence and backbone optimizations, similar to protocols described earlier.28

Sequence optimization was done using Monte-Carlo simulated annealing protocol11; which optimized the total energy of the complex. The all-atom energy function in Rosetta is a linear weighted sum of 12-6 Lennard-Jones potential, the Lazaridis-Karplus implicit solvation model30, an orientation-dependant hydrogen bonding potential31, backbone-dependent rotamer probabilities32, a knowledge-based electrostatic energy33, amino acid probabilities conditioned on φ and ψ space34 and reference energies that approximate the unfolded state energy of an amino acid. 11

**Binding parameters calculation**

The binding energy of a complex was calculated by subtracting the individual Rosetta energies of the chains from the total energy of the complex. Total buried surface area at the interface was similarly calculated by taking the difference in SASAs between the bound and unbound states. The number of buried unsatisfied hydrogen bonding groups was determined by counting the number of polar groups at the interface which were fully buried yet lacked hydrogen bonding partner. The side chain rotamers were left in the same conformations as in the complex when the calculations were made. The interface SASApack scores were reported as the average per-residue SASApack score for those residues within 5 Å of atoms on the opposite chain.
Discrete molecular dynamics

DMD is a flavor of molecular dynamics simulation (MD) approach.\textsuperscript{18, 19, 20} Unlike traditional MD which uses continuous physical force field, in DMD, the interaction between two atoms is described by a simplified step-wise potential. For step-wise potentials, the derivations are zero between the potential steps, and the integrations can be ignored except when the two atoms go across the steps wall (collision events). The space and time of the collision events can be analytically calculated based on previous positions and velocities of the two atoms. The new velocities of the atoms after the collision can also be calculated based on laws of conservation of energy, momentum, and angular momentum. The system evolves by calculation and sorting of further collision events. Compared with traditional MD that is driven by numerical integration over fine time steps, DMD is driven by collision events and is more efficient since it allows larger integration time steps in average. Extra efficiency is also gained by applying fast event sorting and updating algorithms. Although the actual speed up of DMD compared to traditional MD varies from system to system, it can reach 3–10 orders of magnitude.

Experimental methods

Plasmid constructs, gene synthesis and mutagenesis

Kinase dead mutant (K298R) of full length PAK1(PAK1-fl) and PAK1 V127E/S144E (PAK-fl mutant) were cloned in pQE-80 L (Qiagen) vector. Kinase domain (K298R mutant), residues 250-545 from PAK-fl (here referred to as PAK1), was also cloned in pQE-80L vector with an extra sequence (MRGSHHHHHHGSDYDIPTTENLYFQC) in N-terminus. The PAK1 mutants were made by overlap extension using PCR.\textsuperscript{35}
Constructs for protein designs and Spider Roll mutants (from here referred as designed variants) were made as an MBP fusion with a TEV protease cleavage site and cloned in pQE-80L vector so that 6×His-tag remained at N-terminus after the expression. The genes for designed variants were synthesized using a gene synthesis protocol36 where the codons were optimized for bacterial expression.

**Expression and protein purification**

All expressions were carried in BL21(DE3) pLysS cells. The cells were grown up to OD\textsubscript{600} 0.6-0.8 and then induced with 0.3 mM IPTG and further grown at 25 °C for 6 h. Cells were then disrupted using sonication and resulting lysates were cleared by two rounds of centrifugation (18,000 × g) of 20 mins each. The supernatants of PAK1-fl and PAK1 variants were then purified using a prepacked Ni-NTA column (HisTrap, HP, GE Healthcare) followed by an anion exchange step (Source 15Q beads, GE Healthcare) and gel filtration chromatography (Superdex 75 or Superdex 200, GE Healthcare). Designed proteins (referred to as design variants) were purified using Ni-NTA column followed by an overnight proteolysis using TEV protease and then again loading the product on Ni-NTA column. The second Ni-NTA affinity step removed the 6×His-tagged MBP to give design variants (all with an extra GS sequence at N-terminus) as a flow through. Flow through was concentrated and then loaded on gel filtration column (Superdex 75, GE Healthcare). All the design variants eluted at similar elution volume (size corresponding to monomer) and appropriate fractions combined and concentrated (Amicon Ultra, Millipore). Protein concentration was estimated using theoretical molar extinction coefficients and absorbance at 280 nm.

\( U^{15}\text{N} \) and \( U^{15}\text{N},^{13}\text{C} \)-labeled Spider Roll I58A was expressed in M9 minimal media containing \( ^{15}\text{N} \)-ammonium chloride (1 g/l). Glucose was replaced with \( ^{13}\text{C} \)-glucose for \( ^{13}\text{C} \)
labeling. Cells were grown in LB broth up to OD_{600} 0.6 to 0.8 and then spun down by centrifugation at 3000 rpm for 30 mins. The cell pellet was then resuspended in M9 minimal media, left for recovery for 20 mins under 250 rpm shaking and then induced with 0.3 mM IPTG. The expression was carried at 16 °C for 14 h.

**Fluorescence polarization assay**

Binding affinity was measured using fluorescence polarization technique established in our laboratory.\textsuperscript{37} In brief thiol-reactive fluorescent probe Bodipy(507/545)-iodoacetamide (Molecular Probes) was conjugated to design variants at the unique cysteine site. Design variants at a concentration range of 60-250 µM were buffer exchanged into 50 mM Tris-Cl, pH 7.5 using a PD10 desalting column (GE Healthcare) and spiked with 1 mM TCEP. A 20 mM stock solution of Bodipy(507/545-IA) suspended in dimethyl sulfoxide (DMSO) was added drop-by-drop to 3-10 fold molar excess in the designed variants with constant mixing by inverting the tubes. The tubes were wrapped with aluminium foil and conjugation reaction carried for overnight at 4 °C. Next morning β-mercaptoethanol (BME) to a final concentration of 50 mM was added to quench the reaction. The mix was then centrifuged to pellet the unconjugated Bodipy and run over PD10 desalting column (equilibrated with 50 mM Tris-Cl, pH 7.5 and 5 mM BME) to separate the conjugated design variants from free Bodipy. Bodipy labeled design variants were quantified using UV/Vis spectrophotometer (theoretical molar extinction coefficient at 280 nm for protein and molar extinction coefficient of 69,000 M\textsuperscript{-1}cm\textsuperscript{-1} for the probe). A correction factor (Abs\textsubscript{280 nm}/Abs\textsubscript{508 nm}) of 0.03 was used to correct for the absorption of the conjugated probe at 280 nm. Typically probe conjugation efficiency (probe/protein) of 20-90 % was achieved for the design variants.
Fluorescence polarization assays were carried out on a Jobin Yvon Horiba Spex FluoroLog-3 instrument (Jobin Yvon Inc) performed in L-format with the excitation wavelength set at 508 nm and emission wavelength set at 545 nm. Bodipy-design variants (in 50 mM Tris-Cl, pH 7.5, 5 mM BME) at a final concentration of 5-10 µM and volume 180 µl were titrated with PAK1-fl or PAK1 variants (in 50 mM Tris-Cl, pH 7.5, 50 mM NaCl, 5 mM BME). Titrations were performed in 3x3-mm quartz cuvette. Slit width was adjusted to 3.5 nm to give a fluorescence intensity ~3.5 million counts per second. Two polarization readings consisting of 3 averaged measurements were collected for at least 15 concentrations of titrant. Data were averaged and any change in polarization occurring due to addition of buffer was subtracted. Data was then analyzed using a model for single site binding model according to the equation (1) which was incorporated in equation (2) to account for the observed polarization.

\[
[A:B] = \frac{([A_t] + [B_t] + K_d) - \sqrt{([A_t] + [B_t] + K_d)^2 - 4[A_t][B_t]}}{2}
\]

\[
P_{obs} = \frac{(P_{max} - P_0)[A:B]}{[A_t]} + P_0
\]

where \([A:B]\) is the concentration of Bodipy labeled design variants and titrant (PAK1-fl or PAK1 variants) complex formed, \([A_t]\) is the total concentration of Bodipy labeled design variants, \([B_t]\) is the concentration of the titrant, \(K_d\) is the dissociation constant for the interaction, \(P_0\) is the polarization in the absence of titrant, \(P_{max}\) is the maximum polarization observed when all Bodipy labeled design variants are bound to the titrant, and \(P_{obs}\) is the observed polarization at a given concentration of titrant. The data were fit according to equation (2) using non-linear regression with SigmaPlot software to obtain fitted parameters.
for $K_d$, $P_{\text{max}}$ and $P_0$. Delta polarization ($\Delta$Polarization) was calculated using equation (3) and fraction bound was calculated using equation (4).

$$\Delta\text{Polarization} = P_{\text{obs}} - P_{\text{min}} \quad (3)$$

$$\text{Fraction bound} = \frac{P_{\text{obs}} - P_{\text{min}}}{P_{\text{max}} - P_{\text{min}}} \quad (4)$$

where $P_{\text{max}}$ and $P_{\text{min}}$ are maximum and minimum polarization calculated from the fit. Mutants where $P_{\text{obs}}$ was much lower than possible $P_{\text{max}}$, fit was forced to have $P_{\text{max}}$ observed in the fit of Spider Roll-PAK1 interaction (green circles in Figure 3.5). Inability to achieve titrants in very high concentrations put limitations on reaching the saturation level (>10-fold than $K_d$) during a single titration.

**NMR spectroscopy**

As Spider Roll I58A mutant expressed at more than 2-fold higher yield when compared with Spider Roll and since the I58A mutation had a minimal effect on the binding affinity (Figure 3.5 b), we prepared $U^{-15}\text{N}$ and $U^{-15}\text{N}, ^{13}\text{C}$-labeled NMR samples of Spider Roll I58A. For assignment of polypeptide backbone and $^{13}\text{C}\beta$ chemical shifts, NMR spectra were recorded at 288 K on a Varian INOVA 600 spectrometer equipped with a cryogenic probe. 2D-[$^{15}\text{N}, ^1\text{H}$] HSQC (1 h measurement time) was acquired along with four through-bond correlated NMR experiments$^{21}$ that is, HNCA (36 h), HN(CO)CA (36 h), HNCACB (42 h), CACB(CO)NH (42 h) and HNCO (9 h) for sequential resonance assignment. In addition, a 3D $^{15}\text{N}$-resolved [$^1\text{H}, ^1\text{H}$] NOESY spectrum (42 hours) was acquired to confirm sequential resonance assignments and to derive backbone $^1\text{HN}$-$^1\text{HN}$ distance constraints. All spectra were processed and analyzed using the program packages NMRPipe$^{38}$ and XEASY$^{39}$, respectively. Specifically, the titration of Spider Roll with PAK1 was monitored by recording
2D-[$^{15}$N, $^1$H] HSQC spectra at 288 K on a Varian INOVA-700 MHz spectrophotometer equipped with a cryogenic probe. Isotope labeled proteins and kinase domain were buffer exchanged in 20 mM Phosphate buffer, pH 7.0, 50 mM NaCl and 1 mM DTT and 10\% D$_2$O was added. Peak volumes and line-widths in 2D-[$^{15}$N, $^1$H] HSQC were measured by using the ‘peak detection module’ of NMRDraw in NMRPipe. Signal-to-noise ratios (SNR) were calculated by dividing the volume by the noise level measured by the ‘estimate noise module’ of NMRpipe followed by division by the square root of the measurement time. To compare peak volume reductions arising from different degrees of Spider Roll-PAK1 complex formation, the thus obtained SNR per unit time was normalized to the volume measured in the absence of PAK1.

Chemical shifts were deposited in the BioMagResBank (accession code: 16710).\(^{40}\)

**Circular dichroism (CD)**

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans (250-195 nm) were carried at a typical protein concentration of 20 $\mu$M in a 1 mm cuvette. The temperature was maintained at 20°C using a Peltier device. Variable temperature scan was carried between 4-96 °C while measuring the ellipticity at 222 nm. All ellipticity data were corrected with a buffer blank and then converted to mean residue ellipticity.

**Acknowledgments**

Computational services provided in part by RENCI (Renaissance Computing Institute), University of North Carolina at Chapel Hill. The work was supported by DARPA and NIH. Spider Roll protein was selected as a community outreach target by the Northeast Structural Genomics Consortium (http://www.nesg.org; NESG ID OR24). We thank Dr. Ashutosh
Tripathy from UNC Macromolecular Interactions Facility and Dr. Greg Young from UNC Biomolecular NMR Lab for their help during the experiments.

Supplementary information

(a)

![Graph](image1)

(b)

![Graph](image2)
Supplementary figure S3.1. Structural characterization of Spider Roll and mutant scaffold (HYP, PDB ID: 1I2T). (a) Far-UV CD spectrum of 20 µM of Spider Roll compared with the CD spectra of HYP with mutations A15C/G26E/L37E/L38N. (b) CD signal at 222 nm as a function of temperature shows cooperative thermal denaturation. A significant difference in melting temperature ($T_m$) between Spider Roll and HYP is observed. The other designs tested showed similar spectra and cooperative thermal denaturation with differences in $T_m$ and steepness in the slope.

Supplementary figure S3.2. Correlation between the predicted buried surface area and Rosetta binding energy for 43 native complexes (Supplementary table S1). Ten trajectories of DDMI were run to score the near native conformations. The conformation with the lowest binding energy was plotted here against the buried surface area.
Supplementary figure S3.3. 2D-$^{[1\text{H}},^{15\text{N}]}$-HSQC spectra of Spider Roll I58A and backbone resonance assignment. A 600 MHz $^{1\text{H}}$-$^{15\text{N}}$ spectrum of SpiderRoll I58A protein in (20 mM Phosphate, pH 7.0, 50 mM NaCl, 1 mM DTT, 8% D2O), obtained at 288 K. Backbone resonance assignments are labeled. Side chain amides of Asn and Gln are connected by lines. No attempt was made to assign side chain peaks of Asn and Gln residues. Gly6 and Gly56 are folded due to the narrow spectral width of nitrogen dimension.
Supplementary figure S3.4. Comparison of secondary structures determined by chemical shift index (CSI, in blue), CS-Rosetta (in green) and X-ray (in red). For residues which have no CSI output due to lack of assignments, are labeled as ‘×’.

Supplementary figure S3.5. CS-rosetta structure prediction of Spider Roll I58A in an unbound state (green). Crystal structure of scaffold ‘HYP’ (PDB code: 1I2T) (red) and computational model Spider Roll (cyan) are also shown for comparison.
Supplementary figure S3.6. 2D-[\$^{1}\text{H}, \text{^{15}}\text{N}]$-HSQC spectra of Spider Roll with PAK1 or PAK1 L470E. (left) 148 µM Spider Roll I58A with 487 µM PAK1; (right) 150 µM Spider Roll I58A with 490 µM PAK1 L470E mutant. The comparison illustrates that peak volume reduction is significantly less with kinase L470E. The data suggest that mutant interacts with Spider Roll with considerably weaker affinity. The marked residues underwent maximal change in the presence of PAK1.
**Supplementary table S3.1. High resolution PDBs used for evaluation of DDMI protocol**

<table>
<thead>
<tr>
<th>PDB</th>
<th>Resolution (Å)</th>
<th>Partner A / B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A4Y</td>
<td>2.0</td>
<td>Ribonuclease inhibitor / Angiogenin</td>
</tr>
<tr>
<td>1AVA</td>
<td>1.9</td>
<td>Barley alpha-amylase / Barley alpha amyrase (Subtilisin) inhibitor</td>
</tr>
<tr>
<td>1AY7</td>
<td>1.7</td>
<td>Guanyl-specific ribonuclease 5a / Barstar</td>
</tr>
<tr>
<td>1BLX</td>
<td>1.9</td>
<td>Cyclin-dependant kinase 6CDK6 / P19ink4D</td>
</tr>
<tr>
<td>1BRS</td>
<td>2.0</td>
<td>Barnase / Barstar</td>
</tr>
<tr>
<td>1CLV</td>
<td>2.0</td>
<td>Alpha-amylase / Alpha-amylase inhibitor</td>
</tr>
<tr>
<td>1CSE</td>
<td>1.2</td>
<td>Subtilisin carlsberg / Eglin C</td>
</tr>
<tr>
<td>1CXZ</td>
<td>2.2</td>
<td>Human RhoA / Protein kinase PKN</td>
</tr>
<tr>
<td>1D5M</td>
<td>2.0</td>
<td>Enterotoxin type B / HLA Class II Histocompatibility antigen</td>
</tr>
<tr>
<td>1DPJ</td>
<td>1.8</td>
<td>Proteinase A / Proteinase inhibitor IA3 peptide</td>
</tr>
<tr>
<td>1EAY</td>
<td>2.0</td>
<td>CheY/ CheY Binding (P2) Domain</td>
</tr>
<tr>
<td>1EMV</td>
<td>1.7</td>
<td>Colicin E9 DNase/ Immunity protein IM9</td>
</tr>
<tr>
<td>1EUU</td>
<td>1.6</td>
<td>ULP1 Protease / Ubiquitin like protein SMT3</td>
</tr>
<tr>
<td>1F60</td>
<td>1.6</td>
<td>Elongation factor eEF1A / Exchange factor eEF1B alpha</td>
</tr>
<tr>
<td>1FKQ</td>
<td>2.3</td>
<td>RGS domain / GT-G1 chimera unit</td>
</tr>
<tr>
<td>1G4U</td>
<td>2.3</td>
<td>Protein Tyrosine phosphatase SPTP / Ras-related C3 Botulinum Toxin substrate 1</td>
</tr>
<tr>
<td>1GCQ</td>
<td>1.68</td>
<td>VAV SH3 domain / Growth factor Receptor-Bound protein 2 SH3 domain</td>
</tr>
<tr>
<td>1GL4</td>
<td>2.0</td>
<td>Nidogen 1 / PERLECAN IG3</td>
</tr>
<tr>
<td>1HE1</td>
<td>2.0</td>
<td>GAP domain of Exoenzyme S / Ras-related C3 Botulinum Toxin substrate 1</td>
</tr>
<tr>
<td>1HX1</td>
<td>1.9</td>
<td>HSC70 ATPase domain / BAG-family Molecular Chaperone Regulator-1</td>
</tr>
<tr>
<td>1IAR</td>
<td>2.3</td>
<td>Interleukin4/Interleukin 4 receptor alpha</td>
</tr>
<tr>
<td>1IBR</td>
<td>2.3</td>
<td>GTP binding nuclear protein Ran/Importin Beta-1 subunit</td>
</tr>
<tr>
<td>1IM3</td>
<td>2.2</td>
<td>HLA class I histocompatibility antigen A2 / Cytomegalovirus Protein US2</td>
</tr>
<tr>
<td>1IQ5</td>
<td>1.8</td>
<td>Calmodulin / Calmodulin dependent kinase kinase</td>
</tr>
<tr>
<td>1JDH</td>
<td>1.9</td>
<td>Beta catenin / hTcf-4</td>
</tr>
<tr>
<td>1JIW</td>
<td>1.74</td>
<td>Alkaline Mettaloproteinase/ Alkaline Proteinase Inhibitor</td>
</tr>
<tr>
<td>1JLT</td>
<td>1.4</td>
<td>Phospholipase A2/ Phospholipase A2 Inhibitor</td>
</tr>
<tr>
<td>1JTP</td>
<td>1.9</td>
<td>Vh domain fragment of single domain antibody / Lysozyme C</td>
</tr>
<tr>
<td>1PPF</td>
<td>1.8</td>
<td>Human Leukocyte ELastase / Turkey Ovomucoid Inhibitor</td>
</tr>
<tr>
<td>1QMZ</td>
<td>2.2</td>
<td>CDK2/ G2 mytotic-specific cyclin A</td>
</tr>
<tr>
<td>1SGP</td>
<td>1.4</td>
<td>Streptomyces griseus proteinase B / Turkey Ovomucoid inhibitor</td>
</tr>
<tr>
<td>1SVX</td>
<td>2.24</td>
<td>Ankyrin repeat protein/ MBP</td>
</tr>
<tr>
<td>1TX4</td>
<td>1.6</td>
<td>p50 RhoGAP/RhoA</td>
</tr>
<tr>
<td>1U0S</td>
<td>1.9</td>
<td>CheA/CheY</td>
</tr>
<tr>
<td>1UGH</td>
<td>1.9</td>
<td>Uracil DNA Glycosylase/UDG-Inhibitor</td>
</tr>
<tr>
<td>1UZJ</td>
<td>2.1</td>
<td>Designed Colicin E7 Dnase/ Designed Colicin E7 Immunity Protein</td>
</tr>
<tr>
<td>1YCS</td>
<td>2.2</td>
<td>p53/ 53BP2</td>
</tr>
<tr>
<td>2BSR</td>
<td>1.65</td>
<td>Beta lactamase TEM/ Beta lactamase inhibitor protein</td>
</tr>
<tr>
<td>2BKK</td>
<td>2.15</td>
<td>Ankyrin Repeat/ Amino Glycoside phosphotransferase</td>
</tr>
<tr>
<td>2PCC</td>
<td>2.3</td>
<td>ISO-1-Cytochrome C/ Cytochrome C Peroxidase</td>
</tr>
<tr>
<td>2SIC</td>
<td>1.8</td>
<td>Subtilisin BPN' / Streptomyces Subtilisin Inhibitor</td>
</tr>
<tr>
<td>4SGB</td>
<td>2.1</td>
<td>Serine Proteinase B / Potato Inhibitor</td>
</tr>
</tbody>
</table>
Supplementary table S3.2. Backbone assignment of Spider Roll I58A at 288 K.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$^{15}$N</th>
<th>$^{1}$HN</th>
<th>$^{13}$C$_{\alpha}$</th>
<th>$^{13}$C$_{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td></td>
<td></td>
<td>57.76</td>
<td>30.29</td>
</tr>
<tr>
<td>R2</td>
<td>120.66</td>
<td>8.46</td>
<td>58.87</td>
<td>30.26</td>
</tr>
<tr>
<td>Q3</td>
<td>119.27</td>
<td>8.18</td>
<td>58.35</td>
<td>28.30</td>
</tr>
<tr>
<td>A4</td>
<td>122.94</td>
<td>8.13</td>
<td>54.40</td>
<td>18.09</td>
</tr>
<tr>
<td>L5</td>
<td>119.62</td>
<td>8.18</td>
<td>57.78</td>
<td>41.69</td>
</tr>
<tr>
<td>G6</td>
<td>124.42</td>
<td>8.62</td>
<td>48.10</td>
<td>0.0</td>
</tr>
<tr>
<td>E7</td>
<td>119.46</td>
<td>7.85</td>
<td>58.59</td>
<td>29.74</td>
</tr>
<tr>
<td>R8</td>
<td>117.07</td>
<td>7.41</td>
<td>57.89</td>
<td>31.54</td>
</tr>
<tr>
<td>L9</td>
<td>119.44</td>
<td>8.34</td>
<td>57.19</td>
<td>44.07</td>
</tr>
<tr>
<td>Q14</td>
<td></td>
<td></td>
<td>59.39</td>
<td>29.52</td>
</tr>
<tr>
<td>C15</td>
<td>117.66</td>
<td>8.43</td>
<td>57.17</td>
<td>40.26</td>
</tr>
<tr>
<td>K23</td>
<td>123.58</td>
<td>7.99</td>
<td>58.42</td>
<td>32.33</td>
</tr>
<tr>
<td>I24</td>
<td>117.52</td>
<td>8.64</td>
<td>64.74</td>
<td>37.94</td>
</tr>
<tr>
<td>T25</td>
<td>116.74</td>
<td>8.21</td>
<td>68.68</td>
<td>69.50</td>
</tr>
<tr>
<td>E26</td>
<td>118.41</td>
<td>7.67</td>
<td>59.78</td>
<td>29.40</td>
</tr>
<tr>
<td>M27</td>
<td>117.80</td>
<td>7.74</td>
<td>58.95</td>
<td>33.20</td>
</tr>
<tr>
<td>L28</td>
<td>120.87</td>
<td>8.54</td>
<td>57.12</td>
<td>40.64</td>
</tr>
<tr>
<td>L29</td>
<td>115.62</td>
<td>7.74</td>
<td>56.18</td>
<td>40.98</td>
</tr>
<tr>
<td>E30</td>
<td>116.14</td>
<td>7.25</td>
<td>56.49</td>
<td>29.81</td>
</tr>
<tr>
<td>L31</td>
<td>120.56</td>
<td>7.80</td>
<td>54.58</td>
<td>41.64</td>
</tr>
<tr>
<td>S32</td>
<td>114.89</td>
<td>8.49</td>
<td>56.75</td>
<td>62.88</td>
</tr>
<tr>
<td>P33</td>
<td></td>
<td></td>
<td>66.01</td>
<td>31.47</td>
</tr>
<tr>
<td></td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td>Value 4</td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>A34</td>
<td>117.98</td>
<td>8.19</td>
<td>55.00</td>
<td>18.25</td>
</tr>
<tr>
<td>Q35</td>
<td>118.49</td>
<td>7.79</td>
<td>58.54</td>
<td>29.01</td>
</tr>
<tr>
<td>L36</td>
<td>121.22</td>
<td>8.53</td>
<td>58.07</td>
<td>42.20</td>
</tr>
<tr>
<td>E37</td>
<td>118.00</td>
<td>8.44</td>
<td>59.30</td>
<td>28.80</td>
</tr>
<tr>
<td>N38</td>
<td>117.50</td>
<td>7.47</td>
<td>56.22</td>
<td>38.26</td>
</tr>
<tr>
<td>L39</td>
<td>119.12</td>
<td>7.99</td>
<td>57.41</td>
<td>41.12</td>
</tr>
<tr>
<td>L40</td>
<td>115.54</td>
<td>7.85</td>
<td>56.18</td>
<td>41.48</td>
</tr>
<tr>
<td>A41</td>
<td>119.89</td>
<td>7.40</td>
<td>52.56</td>
<td>19.62</td>
</tr>
<tr>
<td>S42</td>
<td>112.71</td>
<td>7.46</td>
<td>55.37</td>
<td>63.26</td>
</tr>
<tr>
<td>E43</td>
<td>123.44</td>
<td>8.23</td>
<td>63.52</td>
<td></td>
</tr>
<tr>
<td>D44</td>
<td></td>
<td></td>
<td>57.24</td>
<td>44.25</td>
</tr>
<tr>
<td>S45</td>
<td>118.68</td>
<td>8.48</td>
<td>62.34</td>
<td>53.30</td>
</tr>
<tr>
<td>L46</td>
<td>123.59</td>
<td>8.29</td>
<td>58.62</td>
<td>39.71</td>
</tr>
<tr>
<td>R47</td>
<td>119.34</td>
<td>8.70</td>
<td>60.04</td>
<td>29.85</td>
</tr>
<tr>
<td>D48</td>
<td>118.81</td>
<td>7.81</td>
<td>57.22</td>
<td>40.00</td>
</tr>
<tr>
<td>Q49</td>
<td>118.38</td>
<td>7.93</td>
<td>57.97</td>
<td>27.64</td>
</tr>
<tr>
<td>V50</td>
<td>122.05</td>
<td>8.83</td>
<td>67.17</td>
<td>31.29</td>
</tr>
<tr>
<td>R51</td>
<td>120.31</td>
<td>7.94</td>
<td>59.78</td>
<td>29.41</td>
</tr>
<tr>
<td>Y52</td>
<td>119.94</td>
<td>8.09</td>
<td>60.07</td>
<td>37.38</td>
</tr>
<tr>
<td>A53</td>
<td>123.36</td>
<td>8.39</td>
<td>54.76</td>
<td>18.30</td>
</tr>
<tr>
<td>M54</td>
<td>115.77</td>
<td>8.30</td>
<td>56.45</td>
<td>31.55</td>
</tr>
<tr>
<td>A55</td>
<td>121.37</td>
<td>7.59</td>
<td>53.95</td>
<td>17.83</td>
</tr>
<tr>
<td>G56</td>
<td>122.69</td>
<td>7.71</td>
<td>45.63</td>
<td></td>
</tr>
<tr>
<td>I57</td>
<td>121.13</td>
<td>7.69</td>
<td>62.77</td>
<td>38.54</td>
</tr>
<tr>
<td>A58</td>
<td>124.16</td>
<td>8.02</td>
<td>52.56</td>
<td>18.55</td>
</tr>
<tr>
<td>F59</td>
<td>118.86</td>
<td>7.87</td>
<td>58.16</td>
<td>39.48</td>
</tr>
<tr>
<td>H60</td>
<td>120.41</td>
<td>8.10</td>
<td>55.35</td>
<td>29.89</td>
</tr>
</tbody>
</table>
**Supplementary table S3.3. Length and composition comparison of the secondary structure elements between X-ray structure and CS-Rosetta structure**

<table>
<thead>
<tr>
<th>Residue Number</th>
<th>Residue name</th>
<th>Secondary structure</th>
<th>Residue Number</th>
<th>Residue name</th>
<th>Secondary structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray structure (1I2T)</td>
<td>CS-Rosetta structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>GLY</td>
<td>Turn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>SER</td>
<td>helix I</td>
<td>1</td>
<td>HIS</td>
<td>helix I</td>
</tr>
<tr>
<td>1009</td>
<td>HIS</td>
<td>turn</td>
<td>1</td>
<td>ARG</td>
<td>helix I</td>
</tr>
<tr>
<td>1010</td>
<td>ARG</td>
<td>helix I</td>
<td>2</td>
<td>ARG</td>
<td>helix I</td>
</tr>
<tr>
<td>1011</td>
<td>GLN</td>
<td>helix I</td>
<td>3</td>
<td>GLN</td>
<td>helix I</td>
</tr>
<tr>
<td>1012</td>
<td>ALA</td>
<td>helix I</td>
<td>4</td>
<td>ALA</td>
<td>helix I</td>
</tr>
<tr>
<td>1013</td>
<td>LEU</td>
<td>helix I</td>
<td>5</td>
<td>LEU</td>
<td>helix I</td>
</tr>
<tr>
<td>1014</td>
<td>GLY</td>
<td>helix I</td>
<td>6</td>
<td>GLY</td>
<td>helix I</td>
</tr>
<tr>
<td>1015</td>
<td>GLU</td>
<td>helix I</td>
<td>7</td>
<td>GLU</td>
<td>helix I</td>
</tr>
<tr>
<td>1016</td>
<td>ARG</td>
<td>helix I</td>
<td>8</td>
<td>ARG</td>
<td>helix I</td>
</tr>
<tr>
<td>1017</td>
<td>LEU</td>
<td>helix I</td>
<td>9</td>
<td>LEU</td>
<td>helix I</td>
</tr>
<tr>
<td>1018</td>
<td>TYR</td>
<td>helix I</td>
<td>10</td>
<td>TYR</td>
<td>helix I</td>
</tr>
<tr>
<td>1019</td>
<td>PRO</td>
<td>helix I</td>
<td>11</td>
<td>PRO</td>
<td>helix I</td>
</tr>
<tr>
<td>1020</td>
<td>ARG</td>
<td>helix I</td>
<td>12</td>
<td>ARG</td>
<td>helix I</td>
</tr>
<tr>
<td>1021</td>
<td>VAL</td>
<td>helix I</td>
<td>13</td>
<td>VAL</td>
<td>helix I</td>
</tr>
<tr>
<td>1022</td>
<td>GLN</td>
<td>helix I</td>
<td>14</td>
<td>GLN</td>
<td>helix I</td>
</tr>
<tr>
<td>1023</td>
<td>ALA</td>
<td>helix I</td>
<td>15</td>
<td>CYS</td>
<td>Bend</td>
</tr>
<tr>
<td>1024</td>
<td>MET</td>
<td>helix I</td>
<td>16</td>
<td>MET</td>
<td>Bend</td>
</tr>
<tr>
<td>1025</td>
<td>GLN</td>
<td>turn</td>
<td>17</td>
<td>GLN</td>
<td>Turn</td>
</tr>
<tr>
<td>1026</td>
<td>PRO</td>
<td>helix II</td>
<td>18</td>
<td>PRO</td>
<td>helix II</td>
</tr>
<tr>
<td>1027</td>
<td>ALA</td>
<td>helix II</td>
<td>19</td>
<td>ALA</td>
<td>helix II</td>
</tr>
<tr>
<td>1028</td>
<td>PHE</td>
<td>helix II</td>
<td>20</td>
<td>PHE</td>
<td>helix II</td>
</tr>
<tr>
<td>1029</td>
<td>ALA</td>
<td>helix II</td>
<td>21</td>
<td>ALA</td>
<td>helix II</td>
</tr>
</tbody>
</table>
1030    SER    helix II    22    SER    helix II
1031    LYS    helix II    23    LYS    helix II
1032    ILE    helix II    24    ILE    helix II
1033    THR    helix II    25    THR    helix II
1034    GLY    helix II    26    GLU    helix II
1035    MET    helix II    27    MET    helix II
1036    LEU    helix II    28    LEU    helix II
1037    LEU    helix II    29    LEU    Turn
1038    GLU    turn    30    GLU    Turn
1039    LEU    bend    31    LEU    Bend
1040    SER    turn    32    SER    Turn
1041    PRO    helix III    33    PRO    helix III
1042    ALA    helix III    34    ALA    helix III
1043    GLN    helix III    35    GLN    helix III
1044    LEU    helix III    36    LEU    helix III
1045    LEU    helix III    37    GLU    helix III
1046    LEU    helix III    38    ASN    helix III
1047    LEU    helix III    39    LEU    helix III
1048    LEU    helix III    40    LEU    helix III
1049    ALA    helix III    41    ALA    helix III
1050    SER    turn    42    SER    Turn
1051    GLU    helix IV    43    GLU    helix IV
1052    ASP    helix IV    44    ASP    helix IV
1053    SER    helix IV    45    SER    helix IV
1054    LEU    helix IV    46    LEU    helix IV
1055    ARG    helix IV    47    ARG    helix IV
1056    ALA    helix IV    48    ASP    helix IV
1057    ARG    helix IV    49    GLN    helix IV
1058    VAL    helix IV    50    VAL    helix IV
1059    ASP    helix IV    51    ARG    helix IV
1060    GLU    helix IV    52    TYR    helix IV
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1061</td>
<td>ALA</td>
<td>helix IV</td>
<td>53</td>
<td>ALA</td>
<td>helix IV</td>
</tr>
<tr>
<td>1062</td>
<td>MET</td>
<td>helix IV</td>
<td>54</td>
<td>MET</td>
<td>helix IV</td>
</tr>
<tr>
<td>1063</td>
<td>GLU</td>
<td>helix IV</td>
<td>55</td>
<td>ALA</td>
<td>helix IV</td>
</tr>
<tr>
<td>1064</td>
<td>LEU</td>
<td>helix IV</td>
<td>56</td>
<td>GLY</td>
<td>helix IV</td>
</tr>
<tr>
<td>1065</td>
<td>ILE</td>
<td>helix IV</td>
<td>57</td>
<td>ILE</td>
<td>helix IV</td>
</tr>
<tr>
<td>1066</td>
<td>ILE</td>
<td>helix IV</td>
<td>58</td>
<td>ALA</td>
<td>coil</td>
</tr>
<tr>
<td>1067</td>
<td>ALA</td>
<td>helix IV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1068</td>
<td>HIS</td>
<td>helix IV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1069</td>
<td>GLY</td>
<td>coil</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


CHAPTER 4

DESIGN OF A METAL-MEDIATED PROTEIN-PROTEIN INTERACTION

Abstract

Metal-mediated protein-protein interactions (PPIs) have several advantages over interactions lacking metal. The most prominent is the enthalpic gain, as metal coordination is stronger than the non-covalent interactions that exist between amino acids. The other advantage is that metal sites at protein interfaces can help align the two partners for a useful PPI and hence contribute significantly to the association rate. Keeping these things in mind we attempted a metal-mediated de novo protein interface design with a more physiologically relevant binding affinity. Three zinc-coordinating residues were designed on ‘scaffold’ proteins, the fourth coordination contributed by a naturally occurring exposed histidine side-chain (His$_T$) on the ‘target’ protein (UbC12 /Ubiquitin). Conformational sampling was achieved by rotating around the Zn-His$_T$ coordination bond and by varying the dihedral angles of His$_T$. Using this 3×1 approach (3 coordinating residues on the ‘scaffold’ and 1 on the ‘target’), a computationally designed protein ‘Spelter’ was created from the bacterial flageller hook-filament junction protein (PDB ID: 2D4X) to interact with ubiquitin as the ‘target’. The success was modest as Spelter showed tight binding with Zn$^{2+}$ ion but did not show visible Zn-mediated interaction with Ubiquitin.
Introduction

Metal ions are indispensable for cellular activities. One-third of all protein structures in the protein data bank (PDB) contain a protein-bound metal ion. These metal ions have both structural as well as catalytic roles to play. One of the most important metal ions is zinc ($\text{Zn}^{2+}$ referred to as Zn from here onwards). Zn is the only metal represented in all classes of enzymes.$^1$ Zn is predominantly 4-coordinated (tetrahedral geometry) for structural role while for catalytic role 5- and 6-coordinations are also observed.$^2$ The geometrical properties of both catalytic and structural Zn are related to very subtle differences in the coordinate bond length and angles that determine the geometry. The residues that most prominently interact with Zn are His and Cys but Asp and Glu are also seen to play both structural and catalytic roles. One major difference between structural and catalytic Zn sites is the ligand preference. For structural roles, Zn is often coordinated to two or more Cys, while in case of catalytic sites, His is the major ligand as it helps Zn to be a better electron acceptor and serve as a Lewis acid.$^2$

A Zn atom site serving a structural role would be an energetic advantage for the protein-protein interface. One of the main contributions comes from huge enthalpy gain from a coordinate bond formed between Zn and ligand atom. These metal-ligand bonds are stronger than other non-covalent bonds formed at the interface. A Zn binding motif at the protein-protein interface has also shown to improve the association rate.$^3$ For example, a non-native Zn binding motif consisting of four ligands was created at the interface of hGH-
hPRLR (human growth hormone – human prolactin receptor). Two ligands from each side of the interface contributed to the Zn coordination. Zn binding only influenced the on-rate of the two proteins confirming that Zn coordination is orientation specific and can contribute to the alignment of the two proteins that interact. The guiding effect of having a Zn site on a protein interface seems to be same as having a knob-hole combination at the interface.

There have also been recent successes with metal-directed self assembly.\textsuperscript{4; 5; 6} Using a four helix bundle protein, Cytochrome \textit{cb}_{562} as a model, Salgado \textit{et al} were able to create a Zn-dependent tetramer of the protein by engineering two dihistidine motifs at the N- and C-termini of the protein.\textsuperscript{4} The monomeric protein had many polar surface residues and the huge interface created by self association was unfavorable due to burial of these polar groups. Zn mediated interactions at the interface was thought to provide enough thermodynamic driving force to favor the formation of tetrameric assembly. In a subsequent study, the group showed that how a small number of metal coordinating mutations on a non-self-associating protein can help overcome the entropic cost of association and promote the optimization of other non-covalent interactions to form more stable architecture which can form even in the absence of metal.\textsuperscript{5} The authors went on to suggest their results could have evolutionary implications.

The Achilles’ heel of metal-mediated interface design is to find the right position for creating a metal binding motif. The challenge is to simultaneously find a position for the Zn-binding motif and create a large buried surface area that is designable. For the previous works of metal-mediated self-assembly of cytochrome \textit{cb}_{562}, the crystal packing interactions were chosen as the starting point for creating metal-interacting motifs and for optimization of non-covalent interactions around them.\textsuperscript{5; 6} In case of hetero-assembly, where two different
proteins are involved in metal-mediated interaction, it becomes more challenging to find two different proteins with a designable crystal packing interface. Apart from that, a metal binding motif across the interface requires designing both sides of the protein interface. An immediate advantage of this approach would be to create orthogonal binding partners of already interacting partners, very similar to a knob-in-to-hole approach which was used to design orthogonal pairs of GoLoco peptide and $\text{G}_i$. Alternatively, a 3×1 approach might allow one to find a naturally occurring single ligand on an undersigned target.

In our effort to create a Zn-mediated interaction, we chose to go for 3×1 approach where 3 Zn coordinating residues were designed on the ‘scaffold’ protein and the last one contributed by the chosen target protein. We hypothesized that it was more common to find a single Zn-ligand at the surface of a protein than having two Zn-ligands capable of simultaneously coordinating a Zn. Zn-mediated protein interface was created three steps: first designing three residues on the ‘scaffold’ protein and then placing the fourth ligand from the ‘target’ at the right orientation with respect to Zn, then rigid-body searching for orientations with compatible surfaces at the interface. At the end the ‘scaffold’ protein was optimized for hydrophobic and polar interactions with the ‘target’.

Results

Model proteins for Zn-mediated de novo design

The criterion for a good ‘target’ protein was to have an exposed His around some potentially hydrophobic residues so that a complementary binding interface can be created on the ‘scaffold’ protein. Ubiquitin has an exposed His at position 68 (H68), with V70, I44 and L8 in close proximity to it. I44 in Ubiquitin has been recognized as a ‘hotspot’ residue for protein-protein interaction. Similarly UbC12 (a NEDD8-conjugating enzyme) has an
exposed His at position 88 (H88) which corresponds to the position F63 in UbCH7 (Figure 4.1). F63 in UbCH7 is a key residue in interaction with the HECT domain of E6AP (PDB ID: 1C4Z). Since these histidines are surrounded by residues known for protein-protein interactions, they are good ‘targets’ for this design strategy.

![Figure 4.1](image)

**Figure 4.1. Selected ‘target’ proteins for Zn-mediated de novo interface design.** (a) Ubiquitin (PDB ID: 1UBQ) shows L8, V70 and I44 residues that involved in protein-protein interaction. Residue H68 (circled) was used as ‘target’ His for Zn-coordination with a 3×1 approach. (b) UbC12 (PDB ID: 1Y8X) showing the exposed His88 that was used for Zn coordination. (c) UbCH7-HECT domain interface (PDB ID: 1C4Z) showing the F63, residue corresponding to H88 in UbC12 (yellow) at the interface against HECT domain (magenta).

Multiple ‘scaffold’ proteins were considered for the de novo interface design. The main criteria for selecting a ‘scaffold’ protein from PDB database was high resolution crystal structure, monomeric, expressible in *E. coli*, absence of disulphide bonds, below 70% sequence identity from any other ‘scaffold’ and of length 80-250 residues. A total of 636 scaffolds were selected (Supplementary table S4.1).

**Design protocol**
A three-step design strategy was used for the 3×1 approach of Zn-mediated protein interface design. In the first step, RosettaMatch\textsuperscript{11} was used to find three Zn ligand positions on all ‘scaffold’ proteins. RosettaMatch was developed earlier for designing novel enzymes.\textsuperscript{12, 13} It uses a transition state model (TS) to search for designable residue sets on a ‘scaffold’ protein that might stabilize the TS and catalyze a reaction. In our case, the TS consisted of a His\textsubscript{T} (histidine from ‘target’) positioned at a right distance and orientation with a Zn atom that is consistent with Zn-coordination geometry (Table 4.1, Figure 4.2). To search for ligand triplets, His/Cys rotamers were placed at all surface positions of the ‘scaffold’ and the TS was placed sequentially. The ensemble of TS positions satisfying the scoring criteria was recorded in a hash table. The hash table was then scanned for overlapping TS positions that satisfied tetrahedral geometry with protein side-chain ligands at the three open Zn-coordination positions. The output file (match) of RosettaMatch protocol consisted of atomic coordinates of these three positions on the ‘scaffold’ plus the TS model. The matches were then filtered for redundancy and quality of geometry based on the sum of squared differences from ideality of the four co-ordinate bond lengths, six tetrahedral angles and dihedral angle ($\chi_3$) that determines the planarity of imidazole ring of His with respect to coordinated Zn. A total of ~1000 good matches were found on 636 ‘scaffold’ protein surfaces.

**Table 4.1. Tetrahedral geometry parameters for Zn-coordination.** The numbers indicate the mean values with standard deviation of parameters observed in protein database.\textsuperscript{2, 11} The numbers were used for imposing constraints during the RosettaMatch search.

<table>
<thead>
<tr>
<th>Atom pair</th>
<th>d (Å)</th>
<th>$\theta_1$ (°)</th>
<th>$\theta_2$ (°)</th>
<th>$\chi_1$</th>
<th>$\chi_2$</th>
<th>$\chi_3$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>His Zn…N</td>
<td>2.05±0.15</td>
<td>109±15</td>
<td>120±15</td>
<td>Free</td>
<td>Free</td>
<td>0/180 (±15)$^{a}$</td>
</tr>
<tr>
<td>Cys Zn…S</td>
<td>2.33±0.10</td>
<td>109±15</td>
<td>109±15</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
</tbody>
</table>

$^{a}$: Determines the coplanarity of Zn and the imidazole ring in His residues. 0° is compatible for Zn…N\textsubscript{δ} and 180° is compatible for Zn…N\textsubscript{ε}. 
Figure 4.2. Tetrahedral geometry parameters for Zn-coordination. TS used for RosettaMatch was HisT-Zn shown in grey color (Zn in grey sphere). The residues in green are other ligands (His and Cys) that complete the tetrahedral geometry. The black (bold) line is the distance parameter. The dotted lines represent angles $\theta_1$ (tetrahedral angle) and $\theta_2$. $\chi_1, \chi_2$ and $\chi_3$ are the dihedral angles shown in arrows. $\chi_3$ determines the coplanarity of the His imidazole ring and the Zn.

In the second step, TS His replaced the ‘target’ His bringing the Zn, ‘target’ and ‘scaffold’ together. The initial conformation represents the RosettaMatch TS-binding conformation grafted into the ‘target’ and ‘scaffold’ and is guaranteed to have a reasonable metal site but not necessarily a complementary interface. A rigid-body conformational search was achieved by rotating the ‘target’ around the Zn-N vector between metal ion and ‘target’ His or by changing the side-chain dihedral angles of His to new angles consistent with Dunbrack rotamers\textsuperscript{14} (Figure 4.3). Any move was accepted or rejected based on Metropolis-Monte Carlo criteria while optimizing the low-resolution centroid score function.\textsuperscript{15}
Figure 4.3. Rigid-body conformational search for Zn-mediated interface design between ‘target’ (yellow) and ‘scaffold’ (cyan) proteins. Rotation around the Zn-N vector and changing the two dihedral angles of the HisT side chain resulted in large changes in rigid-body orientations between the two proteins without disrupting the metal site.

After a specified number of these moves (typically 100 to converge), an all-atom energy function\(^{15}\) was optimized using mutations while accepting or rejecting a move based on Monte Carlo. During the design process, the residues on the two sides of interface that coordinated with Zn were kept fixed and the residues on the scaffold were mutated to any other residue (except Cys) and repacked with side chain minimization. The residues on the ‘target’ were repacked and side chain minimized while keeping the sequence fixed.

Selection criteria of Zn-mediated interface designs

The final selection of the designs was based on the quality of Zn-coordination geometry and the interface design quality. The binding energy density (binding energy / buried surface area) (which excluded the contribution from the Zn coordination), number of unsatisfied hydrogen bonding groups and total number of mutations made in the ‘scaffold’were all considered. In all cases lower number was favored. Six designs were selected for
experiments from as many as 250,000 independent trajectories. The designs with various parametric scores used for selection are shown in Table 4.2. Two designs used an ankyrin repeat protein (PDB ID: 1SVX) as a scaffold, one design used the N-terminal half of *Archaeoglobus fulgidus* xeroderma pigmentous groups B helicase (PDB ID: 2FZ4) as a ‘scaffold’, one design was based on *Plasmodium falciparum* ubiquitin conjugating enzyme (PDB ID: 2ONU) and remaining two designs were based on a fragment from bacterial flagellar hook-filament junction (PDB ID: 2D4X) as ‘scaffold’. A representative design model is shown in Figure 4.4.

Figure 4.4. Computational model of Zn-mediated interaction between Ubiquitin and a designed ‘scaffold’. Ubiquitin (yellow) shows H68 coordinating with Zn and I44 very close to a Trp designed on a scaffold, 2D4X (cyan). Arg-Asp salt bridge is another important interaction in this design. The designed ‘scaffold’ was named Spelter.
Experimental characterization and binding assays with selected designs

The six designs were expressed as 6×His tagged MBP fusion proteins. 2ONU_ubq design (where 2ONU stands for the ‘scaffold’ and ubq for the Ubiquitin ‘target’) aggregated as soon as it was cleaved from the 6×His-MBP protein and was not further characterized. The two designs based on ankyrin repeat (Ank_UbC12_0032 and Ank_UbC12_0097) aggregated in the presence of Zn. 2FZ4_ubq also showed aggregation upon addition of Zn. Other groups have also observed the phenomenon of metal-mediated protein oligomerization and aggregate formation.16 The two designs derived from 2D4X scaffold (2D4X_ubq_0028 or Spelter and 2D4X_ubq_0320) remained soluble when cleaved from 6×His-MBP tag. Note that only Ubiquitin-targeting designs were characterized beyond this point.

The binding interactions between the ‘target’ protein, Ubiquitin and the 2D4X-based designs were measured using isothermal titration calorimetry (ITC) and fluorescence polarization (FP) assays. In case of ITC, Ubiquitin (Sigma) in a buffer having Zn was titrated into either Spelter or 2D4X_ubq_0320 preloaded with same concentration of Zn. In the FP assay, Bodipy-labeled ubiquitin was titrated with 2D4X-based designs and change in polarization measured. In both kinds of experiments no change in signal observed, confirming that Ubiquitin might not be binding with 2D4X-based designs at all.

Zn binding capability of the designs

After finding no metal-mediated Ubiquitin binding, Spelter was used to test whether the designed protein was indeed binding Zn. In order to find the Zn binding affinity of Spelter, ITC experiments were done where Zn was titrated into Spelter. The experiment showed a very tight binding (K_d ~9.5 nM) with a molar ratio of ~1 (Figure 4.5 a). In order to further confirm that the metal is binding at the right site, the two cysteines were reverted back to the
wild-type residues (Spelter_noCys). In an ITC experiment with Spelter_noCys, Zn titration did not show any appreciable heat released or absorbed (Figure 4.5 b).

**Figure 4.5. Isotheraml titration calorimetry (ITC) assays to show Zn binding to Spelter.** (a) 20 µM Spelter was titrated with 250 µM ZnSO$_4$. The titration shows low nanomolar binding. (b) Spelter_noCys (where designed cysteines were reverted back to WT residues) did not show any appreciable change in heat during ITC titrations where 20 µM Spelter_noCys was titrated with 250 µM ZnSO$_4$.

A complementary thermal denaturation CD experiment was done with Spelter in the presence and absence of Zn. We hypothesized that binding of metal would stabilize the protein causing an increase in thermal melting temperature ($T_m$). A far-UV CD scan showed Spelter was helical, consistent with the crystal structure of the scaffold (Supplementary figure S2.1). Thermal denaturation at 222 nm showed cooperative denaturation curve with $T_m \sim$45°C (Figure 4.6 a). When denaturation was carried at a protein:Zn molar ratio of 1:1,
$T_m$ increased to $\sim 52^\circ C$. Further increasing the amount of Zn (protein: Zn = 1:10) resulted in higher $T_m$ ($\sim 55^\circ C$). Apart from the increase in $T_m$, Zn also caused increased cooperativity during thermal denaturation. The effect of Zn on thermal denaturation of protein was not observed for Spelter_noCys (Figure 4.6 b).

**Figure 4.6.** Circular dichroism signal at 222 nm as a function of temperature in the presence and absence of Zn. (a) 15 µM Spelter was used without ZnSO$_4$ and with Spelter:Zn molar ratio of 1:1 and 1:10. Addition of Zn increased the $T_m$ and cooperative nature of denaturation. (b) 15 µM Spelter_noCys was used. Addition of Zn did not show any measurable effect on $T_m$ and unfolding cooperativity.

**Discussions**

This work shows very modest success in designing metal-mediated protein-protein interactions. The challenge in this work was twofold, designing a Zn-binding motif and then creating a Zn-mediated heterodimer. Using RosettaMatch the first hurdle was successfully overcome. Spelter, which had three out of four residues designed onto its fixed-backbone showed a low nanomolar binding to Zn. However, most of the designs aggregated as soon as Zn was added to the protein solution.
Table 4.2. Computational scores of the design models based on which selections were made.

<table>
<thead>
<tr>
<th>Design</th>
<th>Scaffold</th>
<th>Target</th>
<th>Interface size (Å²)</th>
<th>Binding energy density (REU/ Å²)(^a)</th>
<th># of unsatisfied polar groups</th>
<th>SASApack score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ank_UbC12_0032</td>
<td>1SVX</td>
<td>UbC12</td>
<td>1564</td>
<td>-0.011</td>
<td>5</td>
<td>0.73</td>
</tr>
<tr>
<td>Ank_UbC12_0097</td>
<td>1SVX</td>
<td>UbC12</td>
<td>1478</td>
<td>-0.012</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>2FZ4_ubq</td>
<td>2FZ4</td>
<td>Ubq</td>
<td>1704</td>
<td>-0.009</td>
<td>1</td>
<td>-0.48</td>
</tr>
<tr>
<td>2ONU_ubq</td>
<td>2ONU</td>
<td>Ubq</td>
<td>1459</td>
<td>-0.009</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>2D4X_ubq_0028 (Spelter)</td>
<td>2D4X</td>
<td>Ubq</td>
<td>1231</td>
<td>-0.012</td>
<td>4</td>
<td>-0.315</td>
</tr>
<tr>
<td>2D4X_ubq_0320</td>
<td>2D4X</td>
<td>Ubq</td>
<td>1302</td>
<td>-0.012</td>
<td>4</td>
<td>-1.12</td>
</tr>
</tbody>
</table>

\(^a\): The binding energy does not include the contribution from the metal site.

A protein will typically have many potential Zn-coordinating residues on its surface and a designed Zn-binding motif can assist oligomerization. Hence, it seems likely that a Zn-binding motif was created in those cases too: one that bound zinc between monomers.

The failure to show formation of a Zn-mediated heterodimer could have resulted due to the limited enthalpic contribution in the 3×1 approach where only one coordinating bond bridges the interface. To test affinity for a potential ligand at the fourth site, in an ITC experiment, several Zn-binding ligands (2-mercaptoethanol, DTT, pyridine or imidazole) were titrated into Spelter saturated with Zn. The ligands failed to show any binding in those experiments. When Zn coordinates with three residues on Spelter, the fourth ligand is likely to be solvent water to complete the tetrahedral geometry. So when ligands like 2-mercaptoethanol, DTT, pyridine or imidazole were titrated to Spelter, they had to compete...
with 55 M of water. A 2×2 approach where both ‘target’ and ‘scaffold’ contribute equally to Zn-coordination could have more enthalpic contribution across the interface.

There is also a scope of improving the search of conformational space. Backbone minimization during the design was missing in the present protocol. The search space can be also improved by relaxing the geometric parameters (θ₂ and d of the Zn-N vector) using the standard deviation values (Figure 4.2, Table 4.1).

Overall the work still shows encouraging result as far as a rational design of a ligand binder is concerned. A recent publication showed that computational design of a ligand binder is still an unsolved problem.17 In an effort to design a Zn-mediated heterodimer, the present work shows a successful design of a metal binding protein.

**Material and methods**

**Computational design of Zn binding motifs**

Zn binding motifs were designed using RosettaMatch11 which has been developed for enzyme design and is a part of Rosetta molecular modeling program.15; 18 A detailed description of RosettaMatch method can be found elsewhere.11; 12 In summary, for each scaffold first all the designable positions were identified. The residues which had less than 18 neighboring residues at a cutoff distance of 5.5 Å were considered surface residues and designable positions. All Dunbrack rotamers14 of His/Cys were enumerated at each position. The placement was rejected if the side-chain clashes with the protein backbone. A transition state model (TS, the term is a reminiscent from the computational enzyme design for which RosettaMatch was created) consisting of a HisT and a Zn atom (Figure 4.2) was placed near each rotamer so that two of the four sites for a tetrahedral geometry were satisfied. The position of the TS was then recorded in a hash table. The hash table was then examined for
TS positions that were compatible with precalculated rotamer positions in combinations of three such that no two rotamers originated from the same sequence position. An output file ‘match’ consisted of TS coordinates and coordinates of three His/Cys rotamers (with sequence position on the ‘scaffold’) that completed the tetrahedral geometry of a Zn binding motif.

**De novo interface design around a Zn-binding motif**

A ‘match’ consisted of 4 residues (Hist-Zn and 3 ligands with position numbers from the ‘scaffold’). A starting structure was created where 3 ligands were grafted at the specified positions on the ‘scaffold’ and the ‘target’ histidine (H68 on ubiquitin or H88 on UbC12) was aligned with Hist while maintaining the relative orientation of Zn. Huge backbone clashes between the two proteins were common after this stage. Next, the interface design process was done in two stages. In the first stage, a rigid body search in centroid mode was done to identify conformations that relieved the backbone clashes and gave good surface complementarity. The rigid-body search was done by randomly rotating the ‘target’ protein around the Hist-Zn vector or by changing the two dihedral angles of the side chain. After each move the centroid energy score\(^\text{15}\) was calculated and the move was accepted or rejected based on the Metropolis criterion. A ‘Dual Monte Carlo’ approach was used during which the centroid coordinates (centroid pose) as well as all-atom coordinates (all-atom pose) were tracked and stored simultaneously. This allowed Rosetta to maintain fully atomic details of the metal site while scoring based on less-sensitive centroids. Typically 100 rigid body moves resulted in a convergence. During the second stage, the all-atom pose was designed and repacked on the ‘scaffold’ side while only repacking the ‘target’ side. All rotamers (except Cys) was used for the ‘scaffold’ while for the ‘target’, only rotamers of the native
residues were considered. The design process consisted of Monte-Carlo moves where the all-atom energy function\(^\text{15}\) was used to accept or reject moves based on the Metropolis criterion. After the design/repack stage side-chains were also minimized using gradient descent to nearest local minimum.\(^\text{15}\)

**Binding parameters calculation**

The method has been discussed in chapter 3.

**Experimental methods**

**Plasmid constructs**

The ‘target’ genes (ubiquitin and UbC12) were cloned in an in-house vector derived from pQE-80L where a DNA sequence of MBP (maltose binding protein) was added in the 5’ end of the gene. The genes for all design variants were synthesized in-house using overlapping oligos.\(^\text{19}\) The codons used were optimal for bacterial expression.

**Expression and protein purification**

The method has been described in detail in chapters 2 and 3. Protein expression was done in BL21(DE3) pLysS cells. The expressed proteins with 6×His-MBP fusion were purified using a prepacked Ni-NTA column (HisTrap, HP, GE Healthcare). Post cleavage using TEV protease, protein samples were loaded again on same Ni-NTA column and flow through collected. The flow through was concentrated (Amicon Ultra, Millipore) and loaded on a size exclusion column (Superdex 75, GE Healthcare) equilibrated and eluted with MOPS buffer (10 mM MOPS, pH 6.9, 25 mM NaCl and 0.5 mM TCEP).
Isothermal titration calorimetry

Binding affinities were measured using VP-ITC isothermal titration calorimeter (MicroCal, GE Healthcare). For measuring the ubiquitin binding affinity of Spelter, commercially available ubiquitin (Sigma Aldrich) at a concentration of 250 µM in MOPS buffer supplemented with 800 µM ZnSO₄ was titrated in 20 µM Spelter in the same buffer. For measuring the Zn binding affinity of Spelter, 250 µM of ZnSO₄ in MOPS buffer was titrated in 20 µM Spelter exhaustively dialysed in the same buffer. To test Zn-mediated ligand (2-mercaptoethanol, DTT, pyridine, imidazole) binding, a stock of 250 µM ligand was titrated into 20 µM Spelter. A total of 29 titrations of 10 µl each were made. The data were analyzed using Origin50 software and fitted using a model for ‘one-site binding’.

Fluorescence polarization assay

A detailed method for labeling proteins with thiol-reactive fluorescent probe Bodipy(507/545)-iodoacetamide (Molecular Probes) has been discussed in chapters 3 and 4. In ubiquitin, position 28 (Ala) was mutated to Cys to enable labeling with Bodipy probe. UbC12 already has Cys at positions 47, 65, 95, 111 and 181 of which C65 and C111 are exposed and can be labeled with Bodipy probe. Conjugation efficiency was typically very low (5%-10%) which may be due to poor positioning of the site. The experiment was carried in MOPS buffer with 800 µM of ZnSO₄.

Circular dichroism (CD) experiments

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans (260-195 nm) were carried at a typical protein concentration of 15µM in MOPS buffer in a 1 mm cuvette. The temperature was maintained at 20°C using a Peltier device. A variable temperature scan
was carried out between 4-96°C while measuring the ellipticity at 222 nm. ZnSO₄ at final concentration of 15 µM or 150 µM was added wherever appropriate. All ellipticity data were corrected with a buffer blank and then converted to mean residue ellipticity.
Supplementary information

Supplementary figure S4.1. Far–UV CD scan of Spelter and mutant Spelter as a function of Zn concentration. The scan shows no effect of Zn on secondary structure content of the protein.
Supplementary table S4.1. List of scaffolds used for Zn-mediated de novo interface design

<table>
<thead>
<tr>
<th>1a58</th>
<th>1a7s</th>
<th>1aky</th>
<th>1aly</th>
<th>1c81</th>
<th>1cke</th>
<th>1e58</th>
<th>1e5k</th>
<th>1e9m</th>
<th>1ex7</th>
<th>1f4p</th>
<th>1f9y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fdr</td>
<td>1fm4</td>
<td>1g2r</td>
<td>1g8a</td>
<td>1gmi</td>
<td>1gmx</td>
<td>1gnu</td>
<td>1gwm</td>
<td>1h0p</td>
<td>1h1d</td>
<td>1h68</td>
<td>1h6h</td>
</tr>
<tr>
<td>1h7c</td>
<td>1hbk</td>
<td>1he9</td>
<td>1hh8</td>
<td>1hqv</td>
<td>1htj</td>
<td>1hxi</td>
<td>1i1n</td>
<td>1i2a</td>
<td>1i39</td>
<td>1i5g</td>
<td>1i76</td>
</tr>
<tr>
<td>1i8a</td>
<td>1iap</td>
<td>1id0</td>
<td>1ifr</td>
<td>1ijt</td>
<td>1ikt</td>
<td>1im5</td>
<td>1imj</td>
<td>1i02</td>
<td>1ipc</td>
<td>1iqz</td>
<td>1iu9</td>
</tr>
<tr>
<td>1iuh</td>
<td>1iuk</td>
<td>1ixv</td>
<td>1j2a</td>
<td>1j3a</td>
<td>1j84</td>
<td>1jbe</td>
<td>1jffj</td>
<td>1jg1</td>
<td>1jhs</td>
<td>1jvj</td>
<td>1jil</td>
</tr>
<tr>
<td>1jmw</td>
<td>1jos</td>
<td>1jrl</td>
<td>1juv</td>
<td>1jvw</td>
<td>1k1b</td>
<td>1k6k</td>
<td>1k7j</td>
<td>1kgs</td>
<td>1kmq</td>
<td>1kmv</td>
<td></td>
</tr>
<tr>
<td>1kon</td>
<td>1kr7</td>
<td>1ksk</td>
<td>1kw4</td>
<td>1ky3</td>
<td>1kzf</td>
<td>1kzl</td>
<td>1l2h</td>
<td>1l3k</td>
<td>1lb4</td>
<td>1lfp</td>
<td>1lm6</td>
</tr>
<tr>
<td>1lqy</td>
<td>1lu4</td>
<td>1lvg</td>
<td>1m2k</td>
<td>1mb3</td>
<td>1mg4</td>
<td>1mij</td>
<td>1mj4</td>
<td>1m0k</td>
<td>1m0q</td>
<td>1mr3</td>
<td>1mve</td>
</tr>
<tr>
<td>1mvo</td>
<td>1n3y</td>
<td>1nb9</td>
<td>1neg</td>
<td>1nh9</td>
<td>1nio</td>
<td>1nxn</td>
<td>1nqz</td>
<td>1nzw</td>
<td>1nzn</td>
<td>1o08</td>
<td>1o1z</td>
</tr>
<tr>
<td>1o4r</td>
<td>1o6d</td>
<td>1o8v</td>
<td>1o8x</td>
<td>1o9g</td>
<td>1oap</td>
<td>1ocs</td>
<td>1od3</td>
<td>1oh4</td>
<td>1ojq</td>
<td>1opd</td>
<td>1osh</td>
</tr>
<tr>
<td>1p2f</td>
<td>1p4p</td>
<td>1p4x</td>
<td>1p5f</td>
<td>1p5s</td>
<td>1p90</td>
<td>1pa7</td>
<td>1paq</td>
<td>1pbk</td>
<td>1pi1</td>
<td>1pko</td>
<td>1pmh</td>
</tr>
<tr>
<td>1pz4</td>
<td>1q1u</td>
<td>1q7h</td>
<td>1q7r</td>
<td>1q8b</td>
<td>1qcy</td>
<td>1qf9</td>
<td>1qvl</td>
<td>1qw2</td>
<td>1qzm</td>
<td>1r18</td>
<td>1r26</td>
</tr>
<tr>
<td>1r2d</td>
<td>1r2q</td>
<td>1r6j</td>
<td>1r6n</td>
<td>1r9h</td>
<td>1r9w</td>
<td>1rkb</td>
<td>1r1j</td>
<td>1rm8</td>
<td>1roc</td>
<td>1rw1</td>
<td>1rw7</td>
</tr>
<tr>
<td>1rwj</td>
<td>1ryb</td>
<td>1rz4</td>
<td>1s1e</td>
<td>1s21</td>
<td>1s29</td>
<td>1s2o</td>
<td>1s2x</td>
<td>1s35</td>
<td>1s3g</td>
<td>1s3p</td>
<td></td>
</tr>
<tr>
<td>1s68</td>
<td>1s69</td>
<td>1s7z</td>
<td>1s8n</td>
<td>1s9u</td>
<td>1sau</td>
<td>1sux</td>
<td>1svx</td>
<td>1sdi</td>
<td>1sen</td>
<td>1sgw</td>
<td>1sh6</td>
</tr>
<tr>
<td>1sqw</td>
<td>1su0</td>
<td>1svi</td>
<td>1t00</td>
<td>1t3y</td>
<td>1t4w</td>
<td>1t95</td>
<td>1ta0</td>
<td>1tev</td>
<td>1tff</td>
<td>1tke</td>
<td>1tov</td>
</tr>
<tr>
<td>1tq3</td>
<td>1tq5</td>
<td>1tq7</td>
<td>1ts9</td>
<td>1tt8</td>
<td>1ttz</td>
<td>1tuh</td>
<td>1tuv</td>
<td>1tuy</td>
<td>1u02</td>
<td>1u2p</td>
<td>1u3g</td>
</tr>
<tr>
<td>1u61</td>
<td>1u6t</td>
<td>1u7o</td>
<td>1u7u</td>
<td>1u84</td>
<td>1u9c</td>
<td>1u9p</td>
<td>1uhn</td>
<td>1ui0</td>
<td>1ujc</td>
<td>1ulr</td>
<td>1umh</td>
</tr>
<tr>
<td>1unq</td>
<td>1uoh</td>
<td>1uw</td>
<td>1upq</td>
<td>1urr</td>
<td>1ux8</td>
<td>1uxx</td>
<td>1uy4</td>
<td>1uyl</td>
<td>1uz0</td>
<td>1v0a</td>
<td>1v5h</td>
</tr>
<tr>
<td>1v77</td>
<td>1v7r</td>
<td>1vaj</td>
<td>1ve4</td>
<td>1vgl</td>
<td>1vje</td>
<td>1vjx</td>
<td>1vk1</td>
<td>1vk2</td>
<td>1vk6</td>
<td>1vk7</td>
<td>1vkk</td>
</tr>
<tr>
<td>1vku</td>
<td>1vmb</td>
<td>1vr3</td>
<td>1vr8</td>
<td>1vrs</td>
<td>1vyl</td>
<td>1vzw</td>
<td>1w0h</td>
<td>1w0n</td>
<td>1w1d</td>
<td>1w24</td>
<td>1w2l</td>
</tr>
<tr>
<td>1w41</td>
<td>1w4s</td>
<td>1w66</td>
<td>1w8g</td>
<td>1w9w</td>
<td>1wbe</td>
<td>1wd5</td>
<td>1wj9</td>
<td>1wjj</td>
<td>1w18</td>
<td>1wlf</td>
<td>1wlj</td>
</tr>
<tr>
<td>1woj</td>
<td>1wpa</td>
<td>1wqg</td>
<td>1wr2</td>
<td>1wri</td>
<td>1wm</td>
<td>1ws0</td>
<td>1ws6</td>
<td>1wus</td>
<td>1wub</td>
<td>1wv3</td>
<td>1wvh</td>
</tr>
<tr>
<td>1wvn</td>
<td>1wzw</td>
<td>1x0t</td>
<td>1x1r</td>
<td>1x3o</td>
<td>1x6o</td>
<td>1x6z</td>
<td>1x8h</td>
<td>1xbi</td>
<td>1xbn</td>
<td>1xls</td>
<td>1xcl</td>
</tr>
<tr>
<td>1xdz</td>
<td>1xe1</td>
<td>1xj3</td>
<td>1xk5</td>
<td>1xkr</td>
<td>1xmt</td>
<td>1xs5</td>
<td>1xt0</td>
<td>1txq</td>
<td>1xw3</td>
<td>1xww</td>
<td>1y02</td>
</tr>
<tr>
<td>1y63</td>
<td>1y6i</td>
<td>1y81</td>
<td>1y88</td>
<td>1y8c</td>
<td>1y93</td>
<td>1y9p</td>
<td>1y9p</td>
<td>1ye8</td>
<td>1yio</td>
<td>1yn4</td>
<td>1ynb</td>
</tr>
<tr>
<td>1y6p</td>
<td>1ysq</td>
<td>1ylu</td>
<td>1yvd</td>
<td>1yzl</td>
<td>1z06</td>
<td>1z0f</td>
<td>1z0i</td>
<td>1z0m</td>
<td>1z1s</td>
<td>1z2a</td>
<td>1z2m</td>
</tr>
<tr>
<td>1z2u</td>
<td>1z3x</td>
<td>1z4r</td>
<td>1z67</td>
<td>1z6g</td>
<td>1z95</td>
<td>1zat</td>
<td>1zd8</td>
<td>1zd9</td>
<td>1zde</td>
<td>1z18</td>
<td>1zma</td>
</tr>
<tr>
<td>1zn6</td>
<td>1zv9</td>
<td>1zzk</td>
<td>1zoz</td>
<td>2a0j</td>
<td>2ali</td>
<td>2a1v</td>
<td>2a2k</td>
<td>2a4v</td>
<td>2a5j</td>
<td>2a7b</td>
<td>2a7m</td>
</tr>
<tr>
<td>2a8e</td>
<td>2a9o</td>
<td>2a9o</td>
<td>2aak</td>
<td>2aan</td>
<td>2af0</td>
<td>2aj6</td>
<td>2amh</td>
<td>2amy</td>
<td>2ap3</td>
<td>2ar1</td>
<td>2ar5</td>
</tr>
<tr>
<td>2atz</td>
<td>2avk</td>
<td>2avr</td>
<td>2awg</td>
<td>2awk</td>
<td>2azw</td>
<td>2b0c</td>
<td>2b5h</td>
<td>2bbr</td>
<td>2bdv</td>
<td>2bep</td>
<td>2bf0</td>
</tr>
<tr>
<td>2bfw</td>
<td>2bh4</td>
<td>2bk8</td>
<td>2bk9</td>
<td>2bkf</td>
<td>2bl1</td>
<td>2bl7</td>
<td>2bl9</td>
<td>2bn3</td>
<td>2bmd</td>
<td>2bmm</td>
<td>2bmw</td>
</tr>
<tr>
<td>2boo</td>
<td>2brf</td>
<td>2bsn</td>
<td>2bwq</td>
<td>2byo</td>
<td>2bz7</td>
<td>2bzg</td>
<td>2c1f</td>
<td>2c2p</td>
<td>2c3g</td>
<td>2c53</td>
<td>2c60</td>
</tr>
<tr>
<td>2c71</td>
<td>2cal</td>
<td>2cb9</td>
<td>2cb9</td>
<td>2cdn</td>
<td>2etj</td>
<td>2eve</td>
<td>2ew0</td>
<td>2ew1</td>
<td>2ew5</td>
<td>2exu</td>
<td>2eyi</td>
</tr>
<tr>
<td>2f1w</td>
<td>2f21</td>
<td>2f91</td>
<td>2fcf</td>
<td>2fcf</td>
<td>2fcf</td>
<td>2fdj</td>
<td>2fdr</td>
<td>2fe5</td>
<td>2ff7</td>
<td>2ffq</td>
<td>2fg0</td>
</tr>
<tr>
<td>2f1f</td>
<td>2f1w</td>
<td>2f21</td>
<td>2f91</td>
<td>2fcf</td>
<td>2fcf</td>
<td>2fcf</td>
<td>2fdj</td>
<td>2fdr</td>
<td>2fe5</td>
<td>2ff7</td>
<td>2ffq</td>
</tr>
<tr>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
<td>2fup</td>
</tr>
<tr>
<td>2gkg</td>
<td>2gkp</td>
<td>2go2</td>
<td>2gu3</td>
<td>2gui</td>
<td>2gw2</td>
<td>2gwm</td>
<td>2gwr</td>
<td>2h17</td>
<td>2h5p</td>
<td>2haz</td>
<td>2h5b</td>
</tr>
<tr>
<td>2hbw</td>
<td>2hcb</td>
<td>2hcu</td>
<td>2hdo</td>
<td>2hdz</td>
<td>2he4</td>
<td>2hhz</td>
<td>2hia</td>
<td>2hje</td>
<td>2hnx</td>
<td>2hp7</td>
<td>2hpj</td>
</tr>
<tr>
<td>2hpk</td>
<td>2hpk</td>
<td>2hs5</td>
<td>2hsb</td>
<td>2hw4</td>
<td>2hww</td>
<td>2hxm</td>
<td>2hxp</td>
<td>2hzc</td>
<td>2i0m</td>
<td>2i5h</td>
<td>2i5u</td>
</tr>
<tr>
<td>Design</td>
<td>Amino acid sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ank_UbC12_0032</td>
<td>SDLGRKLLEAAARAGQDDEVRLMANGADVNAADKWGTPPLHLAAKSGHLEIVEVLLKHGADVDSAMGYTPLHLWAALMGHLEIVEVLLKHNGADVNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLCLAAHAGYLEIVEVLLKHGADVNAQDKFGKTAFDASIDNGNEDLAELQKL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ank_UbC12_0097</td>
<td>SDLGRKLLEAAARAGQDDEVRLMANGADVNAADKWGTPPLHLAAKSGHLEIVEVLLKHGADVDSAMGYTPLHLWAALMGHLEIVEVLLKHNGADVNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PLCLAAHAGYLEIVEVLLKHGADVNAQDKFGKTAFDASIDNGNEDLAELQKL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2FZ4_ubq</td>
<td>IAEIYERGTIVVKGDAHVPPFAKFSRSGGTYRAKHWYREIEYFESNGIEFVDNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADPIPTPPFDEISLRLRYERMLEYWLYIKWSVPLLPTSGKTHAVAMAAINELST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PTLIVVPTLALAEIQWKERL.GIFGEYVEFGSR1KELKPLTVSTYDSAYVNAEKL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GNRFMLLIFDGDGGLASTSWWQILEMSAPIFVLCLHADNE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2ONU_ubq</td>
<td>SLSRLQCDFTKLMAGYDLRLINHHTIVFLVMFHGPNGTAYEGGIWMVFWALPW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QFPFQSRPHICFMKLLHLPNVHEDTGCVKEDVINQTWTPLVSNVFVEVFLPQLLT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DPNPSDPLNSDASLMLMDKDKNIYEEVKEYVYKYLASKDLWE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D4X_ubq_0028</td>
<td>VVLSQAQAQNSQYALARFATQKVSLVESLVSQVTTAQTQAQEKIVYAGNTLGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Spelter)</td>
<td>DDDRASLATDLQGIRDQLMNLANSTDNGNGYIFAGYKTEAAPFDQATGGYHG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EKSVTQVDSATLIEIGHTGQAQIFNSICACAVPDPGDSEKNFLFVMLTDAAALK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TPVEGNVNEKEAKAAADKTNRLKNSLHNVLEVRWELEWFLLELSAKDGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D4X_ubq_0320</td>
<td>VVLSQAQAQNSQYALARFATQKVSLVESLVSQVTTAQTQAQEKIVYAGNTLGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DDDRASLATDLQGIRDQLMNLANSTDNGNGYIFAGYKTEAAPFDQATGGYHG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EKSVTQVQVDSATMIVIGHTGQAQIFNSICACAVPDPGDSEKNFLFVMLTDAAALK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KTPVEGNVNEKEAKAAADKTNRLKNSLHNVLEVRWELEWFLLELSAKDGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Computational protein design offers a huge challenge in identifying a sequence compatible for a given tertiary structure. The problem becomes more difficult if the backbone is also unknown for the design of a sequence. The problem is addressed by a flexible backbone design approach which iterates between lowest energy sequence prediction for a fixed backbone and lowest energy backbone prediction for a fixed sequence to come up with a minimum energy sequence for a structure. Another challenge in computational protein design is de novo design of a protein interface on a given protein structure for protein-protein interactions.

Redesign of PAK1 auto-inhibitory domain for a biosensor

We used iterative sequence and backbone optimization protocols for redesigning the auto-inhibitory domain (AID) of PAK1 into an affinity reagent. The new secondary structure sequences ranged from 3 residues for a β-hairpin for the design Hairpin to a 20 residue helical sequence for the design AlmostHelix. The goal was to stabilize the AID by building upon the protein core and hence improve the tertiary structure. Our effort to improve the tertiary structure was also done in conjunction with improving the binding affinity of the affinity reagent since the simplest one; WT_IS (residues 83-137) was unfolded and bound with an affinity of only 4 µM. The idea was that a more structured affinity reagent would have lower entropy cost for folding
upon binding and would show tighter affinity. Different approaches but a similar goal were used to come up with three generations of designs, Hairpin, AlmostHelix and PAcKer.

AlmostHelix showed improvement in secondary and tertiary structure but there was no improvement in binding affinity for the PAK1 kinase domain. We concluded that the increase in ‘rigidity’ might have ‘locked’ the binding interface on AlmostHelix in a perturbed conformation. All our effort to crystallize AlmostHelix in an unbound state was not successful because the domain still was not too folded to attain an ordered structure. The chemical shift peaks of 2D-[\textsuperscript{15}N, \textsuperscript{1}H] HSQC of AlmostHelix did not show very well-spread amide peaks confirming that the designed protein did not have a well-defined tertiary structure.

A designed 16-residue helix in the C-terminus of PAcKer did not show any structural advantage over WT_IS. The actual core of WT_IS is between residues 87-128 (number based on PDB ID: 1F3M). The new sequence in the C-terminus made a minimal interaction with the other core residues. What was huge advantage in PAcKer was that the N- and C-termini were brought close to each other, making its insertion in fluorescent proteins (between residues E172 and D173) with ease. CFP-PAcKer bound with a ten-fold improvement in binding affinity to the PAK1 kinase domain. The insertion into a CFP protein where N- and C-termini were constrained in space was analogous to the wild-type AID where the N- and C-termini are constrained by a β-strand swap and peptide-domain interaction (residues 143-146 of AID and the kinase domain) respectively.

CFP-PAcKer when conjugated with a solvent-sensitive dye works as a biosensor in vitro. In future directions, CFP-PAcKer still has a scope of improvement. Long range electrostatic interactions between the newly designed helix and the domain trunc-WT_IS-2 (residues 87-137) can be established by mutating neutral residues to charged residues. The proposed mutations will
also help improve the solubility of the affinity reagent. A dye conjugated affinity reagent often aggregate when microinjected in cells. Preliminary experiments where CFP-PAcKer derived biosensor was microinjected in cells, showed signs of aggregation at the tip of the needle or inside the cells.

Post purification all WT_IS derived designs (except AlmostHelix) showed signs of proteolysis. PAcKer as well as CFP-PAcKer showed proteolysis. Apart from the new sequences added on in AlmostHelix (at N-terminus) and PAcKer (on C-terminus) the major difference is an inclusion of residues 87-99 (of wild-type sequence based on PDB ID: 1F3M) in PAcKer. These residues can be probed more for proteolysis. Some other works from a different group have indicated the range between residues 105 to 118 to be more susceptible to proteolytic cleavage.\footnote{1}

Ultimately the focus is to use the biosensor derived from CFP-PAcKer \textit{in vivo}. The specificity and sensitivity issues will be clearer only after that and will provide more opportunities for improvement of the biosensor.

\textbf{De novo design of a PAK1 binder}

For \textit{de novo} interface design work, we had a modest success in designing specific binders. We developed a DDMI protocol, which uses rigid-body docking, sequence design, and gradient-based minimization of backbone and side chain torsion angles to design low energy interfaces between a ‘scaffold’ and a ‘target’ protein. Using that protocol, from scratch we designed an interface on a ‘scaffold’ protein, hyperplastic disc protein (PDB ID: 1I2T). The designed protein, Spider Roll bound with a dissociation constant of 100 µM to the kinase domain of PAK1. The binding affinity we achieved had low physiological relevance; we think that there were several feats which we achieved.
Spider Roll is a first computational one-sided design. One-sided protein designs have applications as protein therapeutics and affinity reagents for biosensors. Our approach needs only structural information of the ‘target’ with biochemical information which identifies residues on the ‘target’ involved in protein-protein interaction.

We have incorporated docking, design and backbone side chain minimization in our protocol. DDMI can be assumed to do what nature does for designing an interaction between a ‘scaffold’ protein and a ‘target’. Dock stage in our protocol recapitulates nature’s effort to search for complimentary surfaces and burial of potential ‘hotspot’ residues on the target. Design stage of DDMI protocol, recapitulates nature’s effort in coming up with mutations on ‘scaffold’ that can potentially get rid of unwanted steric clashes, and produce useful interactions in the form of hydrogen bonds and salt bridges. Lastly minimization stage recapitulates nature’s induced-fit mechanism and allostery which results in subtle to larger backbone and side chain movements. We believe nature also has to go through a number of iterations of design and minimization to come up with a physiologically relevant interaction.

Our ability to create an alternative binding protein or an affibody for a biologically important protein PAK1 is also an achievement. PAK1 plays role in plethora of functions, ranging from cytoskeletal motility to signal transduction, to apoptosis to cell cycle progression and has shown to be involved in pathological condition and transformation. Some recent reviews have already warranted us with a need for novel engineered protein scaffolds and novel binding protein from nonimmunoglobulin domains. Spider Roll was designed from a nonimmunoglobulin domain which does not have any documented biological relevance with PAK1.
The final notable feat is Spider Roll’s ability to distinguish between a fully active form of PAK1 and ‘closed’ inactive form of PAK1. This will have an important application as an affinity reagent for a biosensor that can be utilized to visualize spatio-temporal dynamics of the activate form of PAK1.

Despite the above feats, we think that we failed in creating a physiologically relevant binding affinity or an affinity that has potential application as therapeutics or affinity reagents. Our best binder Spider Roll still lagged behind most of the binding affinities of the native complexes (in Table 3.1) by 3-5 orders of magnitude or even greater. The Rosetta scores for our designed models could be distinguished from the scores of the native complexes in several areas. First the binding interfaces of the native complexes were distinctly larger than our designed interfaces. Our protocol failed to create/select large interfaces partially due to the nature of the ‘target’ interface and due to the stringency in rejecting higher number of unsatisfied polar groups at the interface. Secondly the native interfaces were much more polar than designed interfaces. There is a tradeoff between hydrogen bonding energy and the solvation penalty. An approximate nature of energy function fails to predict it correctly. Lastly the interfaces were overall poorly packed than the native complexes. The drawback of our DDMI protocol is that the packing quality solely depends on the L-J interaction, which likes to bring the atoms close but avoid any overlap. ‘Clumping’ of atoms can occur in computational designs, where individual atom pairs are more closely spaced than crystal structure.5

Another huge setback observed was in terms of the solubility of the de novo designs. Since we were targeting a hydrophobic patch on PAK1, we expected the designed interface to be hydrophobic. The resulting increase in propensity to aggregate was not recovered by other mutations distant from the interface in some of our designs. A need to introduce solubility score
in the energy function or during the selection criteria will help reduce the failures and give more opportunities to test the computational ability to design a novel interaction.

We also lacked in our focus to work on the association rates of two interacting proteins. NMR studies with Spider Roll and PAK1 confirmed that association rate of the binding was much lower than $10^5 \text{ M}^{-1} \text{s}^{-1}$. The on-rate constants for protein-protein binding can vary dramatically ($1 \times 10^3 \text{ M}^{-1} \text{s}^{-1}$ to $1 \times 10^9 \text{ M}^{-1} \text{s}^{-1}$) but are often near $1 \times 10^6 \text{ M}^{-1} \text{s}^{-1}$.\textsuperscript{6,7} Achieving shape complementarity helps improve the association rate and by choosing a ‘scaffold’ with good structural homology with a natural binder (AID) of PAK1, we addressed the requirement. Another approach is to identify charged residues near the ‘hotspot’ region on the ‘target’ and design oppositely charged residues on the ‘scaffold’.

Metal-mediated protein-protein interactions can also have improved association rates\textsuperscript{8} due to the ‘guiding’ effect to metal site (very similar to knob-hole combination). As a part of future directions of ‘\textit{de novo} protein interface design’ we attempted to design a Zn-mediated interaction. The design, Spelter, selected from a 3×1 approach (one Zn-coordinating ligand from the ‘target’ and another three ligands designed on the ‘scaffold’) for metal-mediated interface design showed no evidence of metal-mediated protein-protein interaction though metal-binding was established.

Overall the successes in this work can be overshadowed by the failures. Understanding the failures can be as important in the learning processes. The knowledge from the experiments and adequately addressing the failures can help improve the reliability of computational protein design.
References


