

**Computational Design of Self-assembling
Proteins and Protein-DNA Nanowires**

Thesis by

Yun Mou

In Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy



CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2014

(Defended May 21, 2014)

© 2014

Yun Mou

All Rights Reserved

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis advisor, Stephen Mayo, in many aspects. He has always been very generous to me in both scientific and personal affairs. Looking back my long Ph.D. life, I could say that I had a very good work-life balance. I learned a lot in this laboratory in a self-directed and thus unstressed way. Given many failed projects that I tried, I have explored much more than I expected before I started my Ph.D. During all these trial-and-error processes, Steve's guidance and patience are very crucial to the final success of my research. Steve has earned my respect in many ways, and I think that is the most important thing of being an advisor.

I would like to thank my committee members, Jacqueline Barton, Bil Clemons and Harry Gray. I feel worryless with such a strong scientific board examining my work. I would especially thank Harry for his kind and fast responses to all my requests for help. His supporting has always made my day.

I would like to thank several people that directly or indirectly helped my thesis project. Jiun-Yann Yu helped me visualize the protein-DNA nanomaterials for the first time using fluorescence microscopy and continued to give me many useful discussion and supports. Theodore Zwang helped me visualize the protein-DNA nanowires using atomic force microscopy. Impressively, with his rich experience, we found what we want to see in the very first experiment, which saved me lots of time. Shing-Jong Huang helped me collect tons of nuclear magnetic resonance data, which finally solved the protein structure that has been waited for so long. I would like to especially thank Lin-Chen Ho, who encouraged me to solve the protein-DNA nanowire structure by X-ray crystallography. I see this structure as my most important experiment in my Ph.D. life.

I would like to thank Marie Ary for her careful and patient proofreading on almost every manuscript of mine. Without her help, I could not imagine how could I finish my candidacy report,

proposition exam, and this thesis. I would also like to thank Rhonda Digiusto for all her efforts on maintaining the laboratory in a super good shape. I could have failed many of the experiments presented in this thesis if Rhonda did not prepare everything ready and correct for me.

In addition, I would like to thank all Caltech members who kindly gave their time to teach me new techniques, help me troubleshoot problems, and discuss science with me. These generous individuals include Jennifer Keeffe, Matthew Moore, Roberto Chica, Timothy Wannier, Bernardo Araujo, Alexandria Berry, Jan Kostecki, Gene Kym, Toni Lee, Alex Nisthal, Ben Allen, Heidi Privett, Christina Vizcarra, Jost Vielmetter, Jens Kaiser, Pavle Nikolovski, Jie Zhou, Julie Hoy, Mike Anaya, Xin Zhang, Kuang Shen, and Chih-Kai Yang.

Besides my research, I also need to thank many good friends accompanying my personal life these years. I would like to thank Jiun-Yann Yu again, being my best friend at Caltech. With hundreds of movies, beers, street jogging, lunch, and dinner we went through together, you have enriched my Ph.D. life to another level. Joe Yeh as another great alcoholic friend of mine, both your high alcohol tolerance and broad knowledge impressed me very much. I would like to thank Ernie's for feeding me everyday. I am grateful for the happy time being with Nicole Wu. I won't forget the encouragement from Yu-Chieh Huang, especially in my most depressing time. I deeply appreciate Chia-Wei Cheng. This journey would never begin and end without your passion and empathy.

Finally, I have to thank my family, Chung-Yuan Mou, Pih Wang, Shin Mou and, Hsiu-Yi Cheng. Needless to say anything specifically.

ABSTRACT

Computational protein design (CPD) is a burgeoning field that uses a physical-chemical or knowledge-based scoring function to create protein variants with new or improved properties. This exciting approach has recently been used to generate proteins with entirely new functions, ones that are not observed in naturally occurring proteins. For example, several enzymes were designed to catalyze reactions that are not in the repertoire of any known natural enzyme. In these designs, novel catalytic activity was built *de novo* (from scratch) into a previously inert protein scaffold. In addition to *de novo* enzyme design, the computational design of protein-protein interactions can also be used to create novel functionality, such as neutralization of influenza. Our goal here was to design a protein that can self-assemble with DNA into nanowires. We used computational tools to homodimerize a transcription factor that binds a specific sequence of double-stranded DNA. We arranged the protein-protein and protein-DNA binding sites so that the self-assembly could occur in a linear fashion to generate nanowires. Upon mixing our designed protein homodimer with the double-stranded DNA, the molecules immediately self-assembled into nanowires. This nanowire topology was confirmed using atomic force microscopy. Co-crystal structure showed that the nanowire is assembled via the desired interactions. To the best of our knowledge, this is the first example of a protein-DNA self-assembly that does not rely on covalent interactions. We anticipate that this new material will stimulate further interest in the development of advanced biomaterials.

TABLE OF CONTENTS

Acknowledgements		iii
Abstract		v
Table of Contents		vi
Tables and Figures		vii
Abbreviations		x
Chapters		
Chapter I	Introduction	1
Chapter II	Using Molecular Dynamics to Predict Domain Swapping of Computationally Designed Protein Variants	14
Chapter III	Computational Design and Experimental Verification of a Symmetric Homodimer	42
Chapter IV	Computational Design of Self-assembling Protein-DNA Nanowires	77
Appendix	Direct Visualization Reveals Dynamics of a Transient Intermediate During Protein Assembly	109

TABLES AND FIGURES

Figure 1-1.	Summary of <i>de novo</i> protein design history using CPD	12
Figure 1-2.	Overall design scheme for a self-assembling protein-DNA nanowire	13
Table 2-1.	Sequences of wild-type ENH, ENH_DsD, E23P, and E24P	32
Table 2-2.	Sequences of X-ray crystallography, dimerization assays, CPD calculations, and MD-derived hinge B factors for each of the proteins tested.	33
Table 2-3.	Data collection and refinement statistics for the crystal structure of ENH_DsD-YFP (PDB: 4NDJ)	34
Table 2-4.	Data collection and refinement statistics for the crystal structure of E23P-YFP (PDB: 4NDK)	35
Figure 2-1.	Size exclusion and polarization fluorescence assays indicate that ENH_DsD-YFP is a high-affinity dimer	36
Figure 2-2.	X-ray crystallography reveals that ENH_DsD-YFP is a domain-swapped dimer	37
Figure 2-3.	B factor analyses from MD simulations for wild-type ENH	38
Figure 2-4.	Polarization fluorescence indicates that E23P-YFP is a high-affinity dimer	39
Figure 2-5.	X-ray crystallography of E23P-YFP shows that a single proline mutation in the hinge recovers the wild-type fold.	40
Figure 2-6.	B factor analyses from MD simulations for the B1 domain protein L	41
Table 3-1.	Library design of the homodimer interface	64
Table 3-2.	Sequences of wild-type ENH, ENH-c2a, NC3-Ncap, and ENH-c2b	65
Table 3-3.	Data collection and refinement statistics for the crystal structure ENH-c2b, with an extra 21-residue tag at N-terminus (PDB: 4NDL)	66
Table 3-4.	NMR statistics for the structure ENH-c2b-Strep (PDB: 2MG4)	67
Figure 3-1.	Steps used to design a C2-symmetrical homodimer	68
Figure 3-2.	SDS-PAGE of purified proteins from soluble fractions ENH-c2b, with an extra 21-residue tag at N-terminus (PDB: 4NDL)	69
Figure 3-3.	Experimental characterizations of the computational library design ENH-c2b, with an extra 21-residue tag at N-terminus (PDB: 4NDL)	70

Figure 3-4.	CD spectroscopy shows that ENH-c2b is a fully refoldable helical protein	71
Figure 3-5.	Characterization of oligomeric state reveals that ENH-c2b is a homodimer with $K_d \sim 130$ nM	72
Figure 3-6.	Crystal structure of ENH-c2b with the long N-terminal tag “MGSSHHHHHSSGLVPRGSHM” (PDB: 4NDL)	73
Figure 3-7.	Solution NMR structure of ENH-c2b (green) superimposed with design model structure (gray), viewed from different orientations	74
Figure 3-8.	NMR spectrum showing intermolecular NOE restraints obtained by $^{12}\text{C}/^{14}\text{N}$ filtered ^{13}C -edited NOESY experiment	75
Figure 3-9.	Two-step design strategy for the functional design of homodimers	76
Table 4-1.	Sequences of wild-type ENH and dualENH	95
Table 4-2.	Data collection and initial refinement statistics for the protein-DNA co-crystal structure	96
Figure 4-1.	Protein-DNA nanomaterial design strategy	97
Figure 4-2.	Design model of irregular bulk protein-DNA nanoparticle	98
Figure 4-3.	Circular dichroism (CD) spectroscopy of dualENH	99
Figure 4-4.	Biophysical characterization of dualENH	100
Figure 4-5.	Fluorescence polarization experiments with dualENH and wild-type ENH	102
Figure 4-6.	Fluorescence microscopy of protein-DNA nanoobjects	103
Figure 4-7.	Microscope imaging experiments	104
Figure 4-8.	Atomic force microscopy of protein-DNA nanowires	105
Figure 4-9.	Co-crystal structure of protein-DNA complex	106
Figure 4-10.	Co-crystal structure of the protein-DNA complex	107
Figure 4-11.	dualENH-DNA binding and nanostructure formation are inhibited at high salt concentrations	108
Figure A-1.	Mapping the interaction interface of the SRP•SR complexes using EPR spectroscopy	135
Figure A-2.	The mobility of spin labels on SR changed upon formation of the early intermediate, stable complex, or both	136

Figure A-3.	Residues I237, Q425, and N426 (green), which changed EPR spectra specifically in the stable complex, are at the conserved motifs (yellow) that mediate N-G domain rearrangement	138
Figure A-4.	Mutations that disrupt the stable complex did not significantly affect the early intermediate	139
Figure A-5.	Fluorescence decay of donor (DACM)-labeled at SRP (C76) under different experimental conditions	140
Figure A-6.	Distance distributions derived from least-square analyses of the TR-FRET data for each FRET pair in the early intermediate (blue), stable complex (red), and early intermediate bound with cargo (green)	141
Figure A-7.	Conformational distribution of the early intermediate is broad, and is restricted by formation of the stable complex or the cargo	142
Figure A-8.	Electrostatic interactions between the N-domains of SRP and SR stabilize the early intermediate and accelerate stable complex assembly	143
Figure A-9.	Change complementarity between SRP and SR's N-domains is essential for the stability of the early intermediate and the kinetics of stable complex assembly	145
Figure A-10.	The 'N' and 'G' groups represent possible conformations within the ensemble of the early intermediate	147
Figure A-11.	Model of free energy landscapes for the protein assembly process	148

ABBREVIATIONS

AFM	atomic force microscopy
bp	base pair
CD	circular dichroism
CPD	computational protein design
DNA	deoxyribonucleic acid
dsDNA	double-stranded deoxyribonucleic acid
DSS	4,4-dimethyl-4-silapentane-1-sulfonate
E_{as}	atomic solvation energy
<i>E. coli</i>	<i>Escherichia coli</i>
ENH	engrailed homeodomain
FFT	fast Fourier transform
FRET	Förster resonance energy transfer
FPLC	fast protein liquid chromatography
HA	hemagglutinin
HSQC	heteronuclear single quantum coherence
IgG	immunoglobulin G
IPTG	isopropyl β -D-1-thiogalactopyranoside
k_{cat}	catalytic constant
K_d	dissociation constant
K_m	Michaelis constant
k_{uncat}	rate constant for an uncatalyzed reaction
LB	Luria Broth
MD	molecular dynamics
MW	molecular weight

NMR	nuclear magnetic resonance
NOE	nuclear Overhauser Effect
NOESY	nuclear Overhauser effect spectroscopy
nt	nucleotide
PCR	polymerase chain reaction
PDB	protein data bank
RMSD	root mean square deviation
RMSF	root mean square fluctuation
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
T_m	melting temperature
YFP	yellow fluorescent protein

Chapter 1

Introduction

Introduction

Computational protein design (CPD) is an automated process that uses a physical-chemical or knowledge-based scoring function to predict amino acid sequences that will fold into a given three-dimensional structure. With current computational resources, over 10^{100} sequences can be virtually screened within a few CPU hours. The ability to efficiently search such enormous sequence space has allowed scientists to design novel proteins that could not be found with other protein engineering approaches (e.g., directed evolution). *De novo* protein design, i.e., designing proteins from scratch, is particularly challenging, and recently many investigators have focused their CPD efforts in this area. *De novo* computational design can be divided into two categories: full-sequence designs and functional designs. These are briefly reviewed below.

Full-sequence designs

Full-sequence design of a zinc finger domain

The first full-sequence design was achieved in 1997 for a 28-amino acid $\beta\beta\alpha$ motif based on the backbone structure of a zinc finger domain (1). A combinatorial library of 1.9×10^{27} possible amino acid sequences was virtually screened and ranked using a physical chemical scoring function. The best (lowest energy) sequence, FSD-1, was chemically synthesized and its structure was confirmed by nuclear magnetic resonance spectroscopy. FSD-1 only shares 21.4% sequence identity to its design parent (Zif268), and a BLAST search showed that it has very low identity to any known protein sequence. This pioneering work demonstrated that the “inverse-folding” problem (finding amino acid sequences that can fold into a given backbone structure) could be solved with CPD, and the resulting sequence may be significantly different from nature’s solution.

Full-sequence design of a novel globular fold

The second breakthrough in full-sequence design was in 2003 with the creation of a novel globular protein fold called Top7 (2). Remarkably, this study did not use an existing protein fold for sequence design. Instead, the authors iterated between sequence space and structure space to

create a novel fold that did not exist in nature. Top7, a 93-residue α/β protein, was obtained via sequence optimization of a combinatorial library containing 10^{186} rotamers. Fifteen cycles of sequence design and backbone relaxation were used to obtain the final Top7 sequence. The Top7 crystal structure is strikingly similar to the design model (backbone RMSD = 1.2 Å). Moreover, its melting temperature (T_m) is above 99 °C, which reflects the accuracy of the force field and the modeling methodology. This work showed that computational design can be used to create new protein folds, ones that have not yet been observed in nature. In addition, the study pointed out the importance of backbone relaxation during the sequence design process.

Functional designs

Enzyme design

The success of full-sequence designs supported the validity of CPD's structure-and energy-based approach for protein engineering. Thus, it seemed reasonable that if an energy-based score could be correlated with protein function, one should also be able to design a protein with desired functionality from scratch. The idea of designing a functional protein *de novo* was first realized in 2001 with the design of a "protozyme," a protein with p-nitrophenyl acetate hydrolysis activity (3). The design hypothesis was based on the transition state theory: a chemical reaction will be accelerated if the catalyst stabilizes the reaction transition state. In this study, the transition state molecule was modeled into the active site pocket of a catalytically inert protein scaffold. The catalytic residues, as well as the surrounding residues, were computationally designed so that the total energy of the system was minimized. The best design, PZD2, had a K_m of ~ 170 μM , a k_{cat} of $\sim 4.6 \times 10^{-4}$ sec^{-1} , and a $k_{\text{cat}}/k_{\text{uncat}}$ of 180. This activity is comparable to that of early catalytic antibodies, but far below that of natural enzymes.

Two *de novo* enzyme designs with improved activities were reported in 2008. Rothlisberger et al. designed enzymes to catalyze the Kemp elimination reaction and obtained rate enhancements of up to 10^5 (4). Application of *in vitro* evolution to the original computational

designs led to even greater enhancements, resulting in >200-fold increase in k_{cat}/K_m (2,600 $\text{M}^{-1}\text{s}^{-1}$). Jiang et al. designed a retro-aldol enzyme that accelerated the reaction by 10^4 (5). These two landmark studies thus demonstrated the ability to create efficient biocatalysts for reactions that could not be catalyzed by naturally occurring enzymes. Recently, Privett et al. used CPD to design a Kemp eliminase (6); extensive directed evolution of this design was then performed (7), leading to an enzyme that accelerated the reaction 6×10^8 -fold, approaching the efficiency of natural enzymes. This dramatic result illustrates that combining the two approaches, CPD and directed evolution, can be a powerful strategy for the generation of novel and highly sophisticated biocatalysts.

Protein dimer design

In addition to enzyme activity, another important property of proteins is their ability to associate and dissociate with other proteins. In 2011, Fleishman et al. designed a protein inhibitor that specifically binds the conserved stem region of influenza hemagglutinin (HA) (8). They computationally docked a library of protein scaffolds to a specified patch on HA, then designed the interface residues to favor these protein-protein interactions. Protein binders with K_d s in the low μM range were created, and subsequent affinity maturation brought these values into the nM range. Crystallographic analysis confirmed that the bound structure closely resembled the computationally designed model. Because the conserved stem region of influenza HA was targeted, the designed proteins exhibited broad neutralizing activity against multiple influenza HA subtypes.

In addition to heterodimer designs, symmetric homodimer designs have also been attempted. In 2011, Stranges et al. exploited β -strand interactions to create a symmetric homodimer with a K_d in the low μM range (9). In 2012, Der et al. designed metal-protein interactions to mediate homodimer formation and obtained a molecule with a K_d in the low nM range (10). The K_d increased to the low μM range when the metal was absent. Importantly, in

both of these studies, X-ray crystallography confirmed that the homodimer structures were nearly identical to their design models.

Protein-ligand design

Another important function of protein is its ability to bind small molecules, such as secondary messengers. In 2013, Tinberg et al. used a computational approach to design ligand binding proteins (11). The proteins exhibited high affinity and selectivity for the steroid digoxigenin (K_d s were in the μ M and sub-nM range before and after affinity maturation, respectively). They determined the X-ray co-crystal structure of two of the designs and showed atomic-level agreement with the corresponding design models.

Protein macrocrystal design

In addition to the interactions between proteins that occur under physiological conditions, protein molecules also interact each other in the crystal form. In 2012, Lanci et al. designed the crystal contacts of a protein so that it would self-assemble in three dimensions to yield macroscopic crystals (12). The crystal structure of this design exhibited sub-Å agreement with the computational model. This work demonstrated that CPD can be used in a new way—to create macroscopic scale materials.

Protein nanomaterial design

In 2012, King et al. designed protein oligomers to form self-assembling nanomaterials (13). The self-assemblies are shaped like cages and mimic the symmetrical structures of viral capsids. Nevertheless, they have novel architectures with specified symmetries. X-ray crystallography of the cage structures revealed atomic-level agreement with the design models. The potential applications of nano-cages include biomedical applications such as drug delivery. This novel work suggests that the possible uses of CPD are just beginning to be explored, and that there are many ways in which it may contribute to materials science.

CPD: progress and possibilities for the future

In the brief review given here, one can clearly see the progression of CPD: (1) full-sequence designs that serve as proof-of-principle studies; (2) functional designs that capture the properties of naturally occurring proteins, such as catalytic activity or protein/ligand binding, and apply them to new reactions or associations; and (3) functional designs that create large assemblies and novel functionalities for materials science applications, such as protein macrocrystals and nanomaterials. This trend (Fig. 1-1) is very exciting because it appears that computational *de novo* design is becoming more versatile. In addition to creating novel proteins with desired biological functions (such as enzymes and binders), we can also endow proteins with new functions that naturally occurring proteins do not have (such as designed crystals and biomaterials). Despite the novelty of macrocrystal design and nano-cage design, these types of materials are not new—they either occur naturally or have been created in the laboratory. One is then compelled to ask, “Can we create a completely new type of material using CPD, and if so, what type of material will this be?” Both DNA alone and protein alone have been used to create self-assembling biomaterials, such as DNA origami (14) and protein fibrils (11). Materials that conjugate DNA and protein using chemical bonds have also been developed for interesting applications, such as immuno-PCR for sensitive antigen detection (15) and programmed positioning of multienzyme cascades (16). However, as far as we are aware, no one has created protein-DNA self-assemblies via purely non-covalent interactions. We therefore decided to see if CPD could be used to design a self-assembling protein-DNA nanomaterial.

Designing a self-assembling protein-DNA nanomaterial

To design a protein-DNA nanomaterial, one must incorporate the protein-DNA interaction plus at least one additional type of interaction (protein-protein or DNA-DNA) into the material. Our plan was to leverage the power of CPD and use protein-DNA and protein-protein interactions to assemble our protein-DNA nanomaterial. As the *de novo* design of protein-DNA interactions has

not yet been done, we decided to exploit the naturally occurring protein-DNA interaction described in this study. Engrailed homeodomain (ENH) is a monomeric protein that binds a specific sequence of double-stranded DNA in a 1:1 fashion (17). This protein has been extensively used for theoretical and computational studies, including many CPD studies. Importantly, it is a three-helix protein that primarily uses helix-3 for DNA binding. Helices 1 and 2 are on the other side of the protein and do not contact the DNA; they form a flat surface that can potentially be designed as a homodimerization interface without disturbing the protein-DNA binding domain. It is therefore an ideal protein for designing a protein-DNA nanomaterial.

Our plan was to homodimerize ENH using computational docking and CPD. We reasoned that the ENH homodimer should be able to bring two fragments of target DNA into close proximity. Furthermore, if we arranged two protein binding sites on opposite sides of a target DNA fragment, “linearization” of the DNA and the designed protein should spontaneously occur when the two components are mixed. We expected the self-assembly to form nanowires with a diameter equal to the length of the DNA fragment. The length of the nanowire would depend on the number of DNA fragments and proteins that associated with each other. The overall design scheme from wild-type ENH to the nanowire is illustrated in Fig. 1-2.

Although the most straightforward way of making this nanowire is to design ENH into a homodimer, the low stability of wild-type ENH ($T_m = 49\text{ }^\circ\text{C}$) makes this design extremely difficult. Attempts to computationally design an ENH homodimer resulted in aggregated proteins when expressed in *Escherichia coli*. We therefore employed an alternative strategy to overcome this problem. In a previous study, CPD was used to create a stabilized variant of ENH called NC3-NCap ($T_m = 88\text{ }^\circ\text{C}$) (18). Using NC3-NCap as the starting scaffold, we designed a variant that was characterized as homodimeric (ENH_DsD). However, X-ray crystallography unexpectedly revealed that the dimer was in fact domain-swapped. We therefore developed a molecular dynamics (MD) protocol to fix the domain-swapping problem. This is described in Chapter 2. We hypothesized that the hinge region involved in domain swapping would have

unusually high flexibility, and that this flexibility would be revealed in a short MD simulation. Our MD protocol predicted that a variant with a single proline mutation, E23P, could revert the domain-swapped ENH_DsD fold back to the wild-type ENH fold. We fused E23P with a yellow fluorescent protein (YFP), and determined its K_d value to be 10 nM using YFP fluorescence. The structure of E23P-YFP was experimentally confirmed by X-ray crystallography; it was homodimeric and retained the wild-type ENH fold. However, the homodimer was formed by the unexpected interaction between E23P and YFP.

In Chapter 3, we describe our work characterizing the biophysical properties of E23P without YFP fused (renamed ENH-c2b). We found that this protein forms a homodimer with a K_d of ~130 nM. Nuclear magnetic resonance spectroscopy showed that its solution structure is similar to that of our design model. The success of the ENH-c2b homodimer design was achieved via a two-step approach in which wild-type ENH was first stabilized, and then this stable variant was dimerized using CPD.

In Chapter 4, we describe how we designed a protein that can self-assemble with a particular DNA to form protein-DNA co-assembling nanowires. The resulting protein, dualENH, bound DNA specifically and still formed a homodimer. Upon mixing dualENH and DNA fragments that contained multiple protein binding sites, a DNA-protein self-assembly formed immediately, as shown by fluorescence microscopy. Atomic force microscopy revealed that the self-assembly formed nanowires if the DNA fragments were designed to have exactly two binding sites on opposite sides of the DNA helix. The width of the nanowire was ~15 nm, which is consistent with the length of the DNA fragment. The length of the nanowire was up to 300 nm. We solved the co-crystal structure of the nanowire. The crystal structure showed that the nanowire was formed via our designed interactions.

In summary, we used CPD to create a new material—self-assembling protein-DNA nanowires. The nanowire self-assembly relies on designing a protein with two functions: homodimerization and DNA binding. These two functions were implemented on two domains

that can work independently and cooperatively. Nature has often evolved multi-domain/function proteins to achieve complicated tasks (e.g., the DNA-binding and activation domains of transcription factors). Recent studies have demonstrated that CPD can be used for the *de novo* design of many individual functions. In this work, we showed that with judicious choices and use of computational design tools, two functions can be incorporated into a single protein such that novel functionality is created. In the future, we expect that CPD will be used in many new applications, perhaps by combining multiple functions in unique ways. The possibilities of *de novo* CPD are just beginning to be explored.

References

1. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278:82-87.
2. Kuhlman B, *et al.* (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302:1364-1368.
3. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. *Proc Natl Acad Sci USA* 98:14274-14279.
4. Röthlisberger D, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190-195.
5. Jiang L, *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387-1391.
6. Privett HK, *et al.* (2012) Iterative approach to computational enzyme design. *Proc Natl Acad Sci USA* 109:3790-3795.
7. Blomberg R, *et al.* (2013) Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* doi:10.1038/nature12623.

8. Fleishman SJ, *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332:816-821.
9. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using β -strand assembly. *Proc Natl Acad Sci USA* 108:20562-20567.
10. Der BS, *et al.* (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J Am Chem Soc* 134:375-385.
11. Tinberg CE, *et al.* (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501:212-216.
12. Lanci CJ, *et al.* (2012) Computational design of a protein crystal. *PNAS* 109:7304-7309.
13. King NP, *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336:1171-1174.
14. Rothmund PWK (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440:297-302.
15. Sano T, Smith CL, Cantor CR (1992) Immuno-Pcr - Very Sensitive Antigen-Detection by Means of Specific Antibody-DNA Conjugates. *Science* 258:120-122.
16. Niemeyer CM, Koehler J, Wuerdemann C (2002) DNA-directed assembly of bienzymic complexes from in vivo biotinylated NAD(P)H : FMN oxidoreductase and luciferase. *Chembiochem* 3:242-245.

17. Fraenkel E, Rould MA, Chambers KA, Pabo CO (1998) Engrailed homeodomain-DNA complex at 2.2 angstrom resolution: A detailed view of the interface and comparison with other engrailed structures. *J Mol Biol* 284:351-361.
18. Marshall SA, Morgan CS, Mayo SL (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol* 316:189-199.

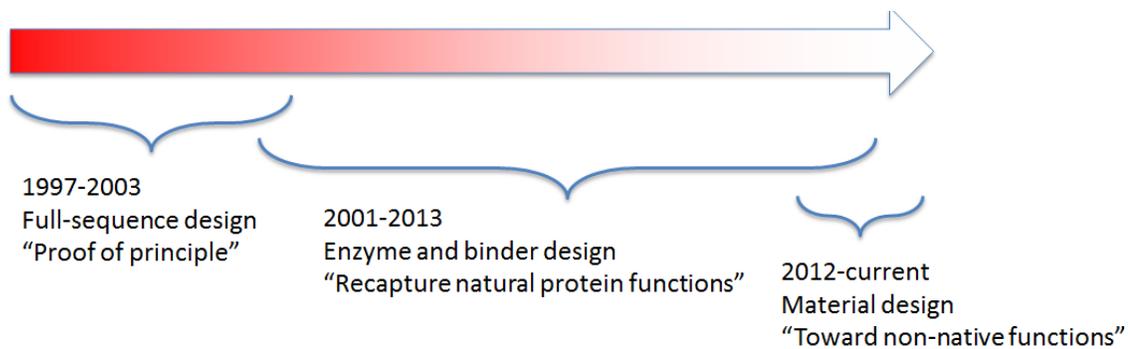


Fig. 1-1. Summary of *de novo* protein design history using CPD.

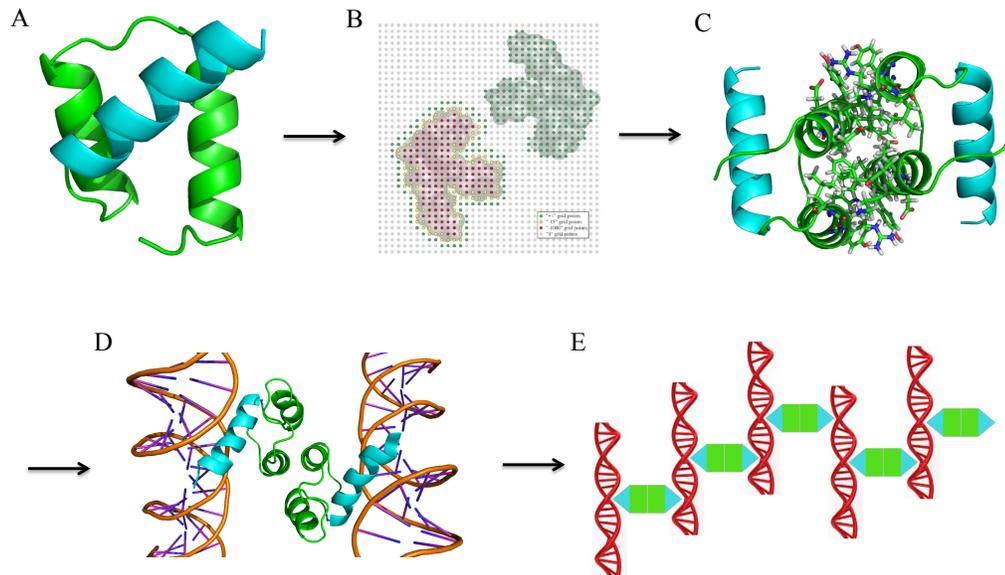


Fig. 1-2. Overall design scheme for a self-assembling protein-DNA nanowire. (A) The protein scaffold, engrailed homeodomain (ENH), is a monomeric protein with three helices. Helices 1 and 2 (green) were chosen for homodimerization. Helix-3 (cyan) is the DNA-binding domain. (B) *In silico* docking was performed to generate a symmetric homodimer configuration using the structure in (A). (C) The interface of the docked homodimer from (B) was computationally designed to create protein-protein affinity. (D) Each component of the designed homodimer specifically binds a dsDNA fragment. The designed homodimer therefore can bring two dsDNA fragments into close proximity. (E) If each dsDNA fragment has exactly two protein binding sites located on opposite sides of the DNA helix, the designed homodimer (green and cyan) and the dsDNA fragment (red) will self-assemble into nanowires.

Chapter 2

Using Molecular Dynamics to Predict Domain Swapping of Computationally Designed Protein Variants

Abstract

In standard implementations of computational protein design (CPD), a positive-design approach is used to predict sequences that will stabilize a given backbone structure. Possible competing states are typically not considered, primarily because appropriate models for them are not available. One of the competing states, the domain-swapped dimer, is especially compelling, because it is often nearly identical to its monomeric counterpart, differing by just a few mutations in the hinge region. Molecular dynamics (MD) simulations provide a computational method to sample different conformational states of a structure. Here, we tested whether MD could be used as a post-design screening tool to identify domain-swapped dimers. We hypothesized that a successful computationally-designed sequence would have backbone dynamics similar to that of the input structure, and that in contrast, domain-swapped dimers would exhibit increased backbone flexibility in the hinge region to accommodate the huge conformational change required for domain swapping. While attempting to engineer a homodimer from the monomeric protein engrailed homeodomain (ENH), we discovered that we had instead generated a domain-swapped dimer (ENH_DsD). We ran MD on these proteins and, as expected, observed increased backbone flexibility in the hinge of the domain-swapped dimer. Two point mutants of ENH_DsD designed to recover the monomeric fold were then tested with our MD protocol. MD predicted that one of these mutants would adopt the monomeric structure, and this was confirmed by X-ray crystallography. Similarly, MD-generated backbone dynamics was found to reflect the domain-swapping tendency of computationally designed variants of the IgG-binding domain of protein L.

Introduction

Computational protein design (CPD) provides *in silico* tools that facilitate the identification of amino acid sequences with specific desired properties. Most CPD algorithms sample an enormous number of amino acid types and side-chain conformations to find the most energy-favored sequences in the context of a single, fixed, main-chain structure (1, 2). Leveraging the speed of modern computers, CPD can effectively reduce the vast sequence space to an affordable number of sequences for experimental examination. CPD is particularly useful when combined with medium to high-throughput experimental screening, and has led to successful designs for a variety of protein engineering problems (3-9). The utility of CPD, however, can be limited in applications where our understanding of the engineering problem is incomplete, or where an appropriate high-throughput experimental screening method does not exist. Another problem that can occur is that the designed sequence doesn't fold into the desired structure, but instead takes on the conformation of a competing state, including unfolded or aggregated states. For example, Fleishman et al. (5) recently showed that only 2 of 88 CPD-designed variants from different protein scaffolds bound to the target molecule, influenza hemagglutinin. A community-wide assessment of this study suggested that many of the failed designs do not adopt the target fold (10). In addition, only half express solubly (11), probably due to poor stability. The ability to predict whether a designed protein sequence will be correctly folded and stable prior to evaluating it experimentally would be extremely beneficial, as it would filter out "poor sequences" so that time-consuming and expensive experimental validations need only be done on sequences that are more likely to have the desired properties.

Although the conformational population of a protein depends on the relative energetic contributions of all possible states, most CPD methods evaluate designed sequences based on only one desired state. Consequently, even though the sequences obtained from these single-state designs may have good CPD scores, this does not ensure that the desired fold dominates the population, because other states may score better than the designed state. Unfortunately, modeling other possible states is not trivial because the conformations of these states are typically unknown.

The stability, specificity, and activity of a protein often depend not only on the protein's structure, but also on its dynamic properties. Altering the dynamics may lead to undesired conformational pathways, such as amyloidogenesis (12). The goal of many protein engineering projects is therefore to maintain the basic structure and dynamics of the protein while improving a desired property (e.g., catalytic activity (13), thermostability (14), substrate specificity (15), ligand binding (16), and molecule transport (17)). However, protein dynamics is typically not modeled in CPD calculations. Fortunately, molecular dynamics (MD) simulations provide a powerful tool for exploring local ensembles of the native state and thus, when incorporated into CPD, allow protein dynamics to be included in the design process. Indeed, Allen et al. showed that MD ensembles could be successfully used for computational multi-state protein design (18).

MD simulations can also serve as a complementary tool to evaluate the dynamic properties of CPD-generated proteins. Both Kiss et al. (19) and Privett et al. (20) used MD simulations as a post-CPD screening method in the *de novo* design of an enzyme to catalyze the Kemp elimination reaction. In these studies, the dynamics of the substrate in the designed active site was monitored using MD. The population of competing states (i.e., bound vs. not bound to substrate) was calculated in the MD trajectories of enzyme variants and used as a filter to identify those likely to exhibit Kemp elimination activity. This approach proved successful and led to the development of the most catalytically efficient computationally designed enzyme for the Kemp elimination to date (20).

Among all the competing alternate states of proteins, domain-swapped dimers are common because they are often nearly identical to their monomeric counterparts (21). Frequently, they differ by only one or two mutations in the hinge region. Apparently, altering these residues can provide the conformational change needed for domain-swapped dimerization while keeping the rest of the protein intact. Studies have shown that mutating these critical residues can affect the domain-swapping tendency significantly (22, 23).

Given that the domain-swapped dimer configuration is usually not modeled explicitly in single-state CPD calculations, it is not surprising that sequences that could assume this fold would be among those predicted, especially if the design involves alteration of loop residues. For example, Baker's group used CPD to design the IgG-binding domain of protein L and inadvertently found that one of the point mutants (G55A) was a weak domain-swapped dimer with a dissociation constant (K_d) of $\sim 30 \mu\text{M}$ (22). Similarly, while attempting to design a homodimer from the monomeric protein engrailed homeodomain (ENH), we generated an even higher affinity dimer ($K_d \sim 40 \text{ nM}$) that also proved to be domain-swapped when examined by X-ray crystallography. Comparison of the crystal structures of ENH and this dimer (named ENH_DsD) suggested that domain swapping might be initiated by opening the loop between two of its helices. We hypothesized that ENH_DsD's ability to make this dramatic move would be reflected by unusually high backbone flexibility along the loop in the monomeric state, and that the wild-type protein (ENH), which does not adopt the domain-swapped configuration, would have significantly lower flexibility in this loop. We anticipated that these differences in loop dynamics might be observable in MD simulations of the two proteins, and set out to explore this possibility. As expected, short 20 ns MD simulations revealed greater flexibility in this loop for ENH_DsD than for the wild type. Similarly, we reasoned that any mutations to ENH_DsD that caused the protein to recover the monomeric state would also be reflected in wild-type loop dynamics. Again, this proved to be the case — an ENH_DsD point mutant that showed wild-type loop dynamics was confirmed by X-ray crystallography to assume the wild-type monomeric fold. To determine the general applicability of our MD protocol, we also investigated two domain-swapped dimer mutants of the IgG-binding domain of protein L, and found that their hinge dynamics correlated with the strength of domain-swapped dimerization.

Results and Discussion

De Novo Homodimer Design Produced a Domain-Swapped Dimer (ENH_DsD). Our original goal was to use computational tools to engineer a *de novo* homodimer from a monomeric protein. We docked two *Drosophila melanogaster* engrailed homeodomain monomers (PDB ID: 1ENH) (24) to form an *in silico* C2-symmetry homodimer (25), then used CPD to redesign the protein-protein interface to create affinity between the two identical chains. The oligomeric state(s) of the designed variants was determined using size exclusion chromatography and fluorescence polarization techniques. To assist in these assays, a yellow fluorescent protein (YFP) (26) was fused to the C terminus of each of the designed sequences. Size exclusion chromatography revealed that one of the predicted sequences, ENH_DsD, elutes primarily as a dimer, with some contribution from higher oligomers (Fig. 2-1A). Fluorescence polarization was determined using the Förster resonance energy transfer (FRET) assay, which yielded a K_d of ~40 nM (Fig. 2-1B). The decreased polarization at higher concentrations is caused by the dimerization of ENH_DsD-YFP, in which two nearby YFPs transfer energy to each other (homo-FRET). Note that these assays cannot distinguish between a regular dimer (where each chain retains the wild-type monomeric fold) and a domain-swapped dimer. The only definitive way to make this distinction is to solve the structure, so we performed X-ray crystallography on ENH_DsD-YFP and resolved the structure to 1.85 Å (Fig. 2-2A). Analysis revealed that a domain-swapped dimer formed between two ENH_DsD sequences (Fig. 2-2B). It appears that the hinge-loop between the first and second helices in the wild-type structure flipped over and coiled up. In the domain-swapped conformation, the first helix, the hinge, and the second helix thus form a single long helix (Fig. 2-2C). The rest of the structure essentially remains intact; i.e., excluding the hinge, superposition of ENH_DsD and wild-type engrailed gives a C_α RMSD of 0.57 Å (Fig. 2-2B). Note that ENH_DsD shares only 49% sequence identity with wild-type ENH (25 out of 51 residues).

Molecular Dynamics Suggests Increased Hinge Flexibility Is Associated with ENH_DsD's Domain-Swapping Capability. Given our goal of engineering a regular homodimer, we were displeased to find that ENH_DsD is a domain-swapped dimer. Unfortunately, solving the structure of all promising dimer designs in order to verify that they retain the wild-type fold would be extremely time-consuming. We hoped to circumvent this laborious process by developing a computational method that could predict whether designed variants are domain-swapped or not.

In the case of ENH_DsD, the hinge between helices 1 and 2 must be highly flexible to allow it to flip over $\sim 90^\circ$ and coil up to assume the domain-swapped dimer configuration. We hypothesized that flexibility in the hinge would be observable with MD, and that an MD simulation of the ENH_DsD sequence threaded onto the wild-type monomer structure would show unusually high hinge dynamics. However, any loop is intrinsically more flexible than secondary structure regions whether it is involved in domain swapping or not. A reference was therefore needed to define “unusually” high dynamics. We used the wild-type protein as our reference, because it is known to be a stable monomer that does not exhibit dramatic conformational changes. We ran 20 ns explicit water MD simulations on the wild-type protein and on the ENH_DsD sequence threaded onto the wild-type backbone structure, and for each trajectory calculated the backbone root mean square fluctuation (RMSF) for each C_α over the whole trajectory. Three trajectories were run for each sequence, and the average B factors were calculated and used as our metric of structural flexibility. As seen in Fig. 3A, nearly all the residues in the wild-type protein have very low B factors, except those close to the N and C termini. The B factors for the ENH_DsD residues are similar to wild type, except for those in the hinge, particularly residues 23 and 24, which are significantly higher (Fig. 2-3B). Unusually high flexibility in this hinge may thus be associated with ENH_DsD's domain-swapping capability. It is entirely possible, however, that this enhanced hinge flexibility could instead reflect other physical phenomena, such as switching between multiple loop conformations.

ENH_DsD Mutant Designed to Recover Monomeric Fold Reverts to Wild-Type Hinge Flexibility. As mentioned above, if ENH_DsD's domain-swapping capability is associated with enhanced flexibility in the hinge, we reasoned that mutations to ENH_DsD that cause the protein to recover the monomeric fold would also change its hinge dynamics back to wild-type. An obvious strategy for reverting ENH_DsD back to the wild-type structure is to put one or more prolines in the hinge. Steric conflict from the pyrrolidine ring in proline precludes residues directly N-terminal to it from adopting a helical structure. By discouraging helix formation in the hinge of the domain swapper, we hoped to shift the equilibrium so that the monomeric fold would again become the dominant state.

Since prolines often occupy the first or second residue of α -helices (27), we substituted a proline into these positions on helix 2 of ENH_DsD to make two point mutants (E23P and E24P) and examined the dynamics of these sequences using the MD protocol described above. As seen in Fig. 2-3B and C, the B factors of the hinge residues in E23P are much lower than those in ENH_DsD; in fact, although the error bars are larger, E23P's B factors are very similar to those of the wild type at all positions (Fig. 2-3A and C). These results suggest that the E23P sequence is dynamically stable in the wild-type conformation and less likely to switch to other conformations. In contrast, E24P resembles ENH_DsD in that it still has a very high B factor region that corresponds to the hinge-loop (Fig. 2-3B and D). Thus, the hinge dynamics obtained from MD show that E23P is wild-type-like, and thus probably a regular dimer, whereas E24P resembles ENH_DsD and is therefore more likely to form other competing states. The sequences of wild-type ENH, ENH_DsD, E23P, and E24P are listed in Table 2-1.

Biophysical and Structural Analyses Support MD Predictions. We constructed the E23P-YFP and E24P-YFP proteins and examined their dimerization properties using the homo-FRET assay. E23P-YFP appears to be a strong dimer ($K_d \sim 10$ nM) (Fig. 2-4), but E24P-YFP shows no detectable dimerization in the low μ M range (anisotropy >300 mA). The fact that E23P-YFP and E24P-YFP

differ in their oligomeric states is consistent with the MD data, and suggests that these two proteins adopt different conformations.

To confirm our MD prediction that E23P is a regular dimer and not a domain-swapped dimer, we determined the structure of E23P-YFP to 2.3 Å using X-ray crystallography (Fig. 2-5A). Superposition of the wild-type (ENH) and E23P-YFP structures (Fig. 2-5B) clearly shows that the single proline mutation in E23P rescues the wild-type fold. The C_{α} RMSD for E23P vs. wild type is 0.36 Å for the whole chain and 0.62 Å for the hinge-loop. The X-ray structure thus confirms that the dimerization observed by homo-FRET results from a regular dimer. Interestingly, the dimer is not formed as modeled in the initial CPD homodimer designs. Instead of a single interface being formed by the two designed surfaces of the first and second helices, two interfaces are formed between each of the designed surfaces and a patch on YFP (Fig. 2-5A). This unexpected finding emphasizes that spurious results can occur when unwanted conformations are not specifically excluded by incorporating negative design into CPD.

Protein L Mutants Test Applicability of MD Protocol. We also investigated two mutants of the B1 domain of protein L from *Peptostreptococcus magnus* generated by Baker's group (22, 23). O'Neill et al. showed that the point mutant G55A is a weak domain-swapped dimer ($K_d \sim 30 \mu\text{M}$) (22). CPD was then used to stabilize this domain-swapped dimer structure, creating a triple mutant obligate dimer ($K_d \sim 700 \text{ pM}$) (A52V/N53P/G55A) (23). We decided to take advantage of the large difference in dissociation constants exhibited by these mutants to test the applicability of our MD protocol. We therefore applied our MD method to three protein L sequences (wild type, G55A, and the triple mutant) to determine whether the backbone dynamics of the hinge reflects the protein's domain-swapping tendency.

Each of the three sequences was threaded onto the wild-type monomeric structure (PDB ID: 1HZ5) (28) and the same MD protocol developed above was applied. Compared to wild-type ENH, wild-type protein L has a more flexible hinge region (Fig. 2-5A). This high dynamics is an intrinsic

property of the hinge-loop and does not necessarily reflect a domain-swapping capability. We used the wild-type again as our baseline for determining unusually high dynamics. As seen in Fig. 2-5B, the hinge dynamics of the weak domain-swapped dimer G55A are not significantly different from wild type. This result is not surprising, as G55A is a fairly stable monomer (2.6 kcal/mol) at concentrations under 10 μ M. In contrast, the triple mutant shows a rather large change in hinge dynamics (Fig. 2-5C; B factor increases significantly at position 53), reflecting its strong domain-swapping tendency ($K_d \sim 700$ pM).

Table 2-2 summarizes the hinge B factors for each of the proteins tested, and compares these data with results obtained from X-ray crystallography, dimerization assays, and CPD calculations. First, note that domain swapping does not correlate with CPD energies. The domain-swapped dimer ENH_DsD has a lower CPD energy (in the context of the monomer structure) than the regular dimer E23P, which prefers the monomeric state. This poor correlation is a direct result of failing to consider competing states in CPD. On the other hand, of the three experimentally validated domain-swapped proteins, the two stronger dimers (ENH_DsD and A52V/D53P/G55A) showed more flexibility in the hinge, as indicated by their MD-derived B factors. It appears that our MD method can identify domain-swapped dimers with K_d values in the nanomolar range or lower. However, this is a preliminary conclusion based on only three cases, and should be confirmed with further studies. The mechanisms underlying the dynamics of domain swapping are not well understood and may vary for different proteins.

Molecular Dynamics as a Post-CPD Screening Method. CPD is a powerful tool for predicting sequences that may solve a given protein engineering problem. However, due to imperfections in CPD (e.g., inaccurate scoring function, fixed backbone approximation, discrete rotamer approximation, implicit water modeling, and lack of dynamics information) typically only a small portion of the top ranked sequences actually satisfy the design goals. The utility of CPD is thus critically dependent not only on the design strategy, but also on the experimental screening method

used to identify hits from the sequences predicted. Higher-throughput screens are obviously preferred, if available, as they have the distinct advantage of allowing larger libraries to be tested.

MD simulations can serve as a complement to CPD and can help make up for some of its limitations. The continuous structure space of MD gives a much more accurate description of protein softness. The explicit inclusion of water is also a key feature for accurate protein modeling, especially for surface/pocket designs, such as loop designs, binder designs, or enzyme designs. More importantly, the trajectories provided by MD simulations provide time-related information that can be very informative. Several groups have thus used MD simulations as a post-CPD filter method for various engineering purposes. For example, Kiss et al. (19) used MD to evaluate and re-rank variants obtained from *de novo* enzyme designs for the Kemp elimination reaction. Multiple analyses were developed as criteria for filtering the most promising designs. Among them, the dynamics of the system, instead of a single snapshot or an averaged structure, was used to evaluate enzyme activity. Key contacts between the protein sidechains and the substrate were monitored throughout the trajectory. Also, water accessibility in the active site pocket was analyzed, as water molecules are abundant and compete with the substrate. Overall, active designs could be clearly distinguished from inactive ones in 20 of the 23 cases tested. This is an impressive improvement over the success rate of CPD alone (8/59). However, as in our studies, their MD protocol was less successful in identifying marginally active enzymes. Privett et al. used a similar MD protocol to suggest a point mutant that resulted in a 3-fold improvement in Kemp elimination activity (20).

Liang et al. also reported studies in which MD simulations were used to validate designs (29). They showed that the stability of a designed complex containing a computationally-grafted binding epitope could be visualized by the MD trajectory. Although their root means square analysis requires a much longer MD simulation time (~400 ns), it is conceptually similar to our RMSF protocol. Again, this study emphasizes the importance of protein dynamics—whereas MD analysis was able to distinguish stable protein complexes from unstable ones, monomer-dimer ΔG values calculated from static structures could not. Although their engineering goals were different,

the similarity between Liang's and our methods indicates that analysis of MD-generated dynamics information can be a sensitive metric for evaluating the effects of even small perturbations in proteins. We expect that our MD protocol, which monitors backbone dynamics, will be especially useful, given that for most CPD engineering projects, the predicted variants must assume the correct fold in order to meet the design goals.

The MD protocol proposed here can efficiently examine the fast backbone dynamics of proteins. A 20 ns trajectory of a small 50–60-residue protein such as ENH or protein L can be run in one day on a modern 8-CPU node with a parallelizable MD package such as GROMACS (30). Our protocol, however, is only applicable for fast dynamics (i.e., in the nanosecond range). The dynamics of many other events (e.g., protein-protein association or dissociation) occur on much longer time scales (from microseconds to seconds). In these cases, either a longer trajectory, such as that employed in Liang's work, or a different, more efficient type of MD (e.g., coarse-grained models or steered MD) might be required to identify good designs. Significant effort has been applied to generate ultra-long MD simulations so that events with slower dynamics, such as protein folding, can be studied (31). We anticipate that as MD techniques continue to improve, the utility of MD as a post-CPD screening tool will expand and become more prevalent.

Conclusions

In this work, we present a simple MD protocol for evaluating the domain-swapping tendency of CPD-designed protein variants. The protocol consists of a 20 ns MD simulation of the designed sequence, followed by RMSF analyses of the backbone atoms. By applying this protocol to the domain-swapped dimer ENH_DsD, we found unusually high flexibility in the hinge-loop compared to that of the monomeric wild-type protein. To recover the wild-type fold, two mutants, E23P and E24P, were made and examined using our MD protocol. E23P exhibited wild-type-like backbone dynamics and was therefore predicted to revert to the wild-type fold. Resolution of the structure by X-ray crystallography confirmed the MD prediction. We tested the applicability of our method on

CPD–designed variants of the B1 domain of protein L, and similarly found that domain-swapping tendency correlated with flexibility in the hinge region. We anticipate that our MD protocol can be used as an in silico post-CPD filter, which may circumvent the need for time-consuming structural studies and will be most useful when experimental screening techniques are not adequate.

Materials and Methods

Construct Preparation, Expression, and Purification. Oligonucleotides (Integrated DNA Technologies) containing ~20 bp overlapping segments were assembled via a modified Stemmer polymerase chain reaction (PCR) method using KOD Hot Start Polymerase (Novagen) to generate the full-length designed sequence ENH_DsD (Table S1). The PCR product and YFP gene (26) with a C-terminal His₆ tag at the 3' end were fused using overlap extension PCR. The ENH_DsD-YFP gene was then cloned into pET-11a using standard digestion/ligation methods. The E23P and E24P mutants were created by standard quick-change protocols. The plasmids were transformed into *Escherichia coli* BL21(DE3) cells; colonies were picked, and the plasmids minipreped and sequenced. Sequence-verified constructs were expressed in standard Luria Broth at 37°C using 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 3 h. The cells were centrifuge harvested and sonication lysed. The target protein was purified by Ni²⁺-NTA (Qiagen) column.

Size-Exclusion Chromatography. Size-exclusion chromatography was carried out at room temperature using an analytical Superdex-75 column (Amersham Pharmacia). After affinity column purification, each sample was loaded on an ÄKTA FPLC system with 0.5 mL sample volume, and run at 0.5 mL/min flow rate with running buffer (100 mM NaCl and 20 mM Tris-HCl, pH 8.0). Absorbance at A515 (absorbance peak for YFP) was tracked for protein elution. Typically, >10 mg protein/L cell culture was purified for each construct.

Polarization Fluorescence Assay. The polarization fluorescence was measured at room temperature with a Fluorolog-3 spectrofluorometer (HORIBA). ENH_DsD-YFP was serially diluted in buffer containing 100 mM NaCl and 20 mM TrisHCl at pH 8.0. The fluorescence

anisotropy was measured for each sample, and the G-factor was determined individually. The data were analyzed according to a simple monomer-dimer equilibrium model and fit with KaleidaGraph software. The anisotropy values (mA) for the completely monomeric and dimeric states were fit to be 260 and 330, respectively.

Crystallography. The ENH_DsD-YFP crystal was grown in 0.8 M monosodium phosphate, 1.2 M dipotassium phosphate, and 0.1 M sodium acetate at pH 4.5 using hanging-drop diffusion. The E23P-YFP crystal was grown in 0.1 M sodium chloride, 0.1 M 12% v/v/ 2-propanol, and 0.1 M sodium acetate at pH 4.6 using hanging-drop diffusion. Crystals were flash frozen in glycerol cryoprotectant and shipped to beamline 12-2 at Stanford Synchrotron Radiation Lightsource. Phases were obtained through molecular replacement using YFP as a model (PDB:1MYW) (26). Following molecular replacement, the ENH_DsD and E23P residues were built manually into the electron density map using COOT (32), respectively. Further refinement was done by PHENIX (33). Final coordinates were deposited in the Protein Data Bank with the codes 4NDJ (ENH_DsD-YFP) and 4NDK (E23P-YFP). Data collection and refinement statistics are listed in Tables 2-3 and 2-4.

MD Simulations. The input structures for the MD simulations were prepared as follows: 1ENH and 1HZ5 were used as the template backbone structures for the engrailed homeodomain and IgG-binding domain of protein L sequences, respectively. Sequences were first threaded onto the corresponding backbone structure and side-chain repacking optimization was applied. The structure was then input to GROMACS 4.5.5 for energy minimization (GROMOS 43a1) with an explicit water box under periodic boundary conditions (30). If there were any extra charges on the protein, they were neutralized by adding sodium or chloride ions. After energy minimization converged to $F_{\text{max}} < 1000$ kJ/mol, a 20 ps position-restrained MD simulation was run for water relaxation at 300 K with 2 fs steps. Finally, a 20 ns unrestrained MD was run at 300 K under NPT conditions. The RMSF of C_{α} atoms was analyzed by `g_rmsf` built in GROMACS for the whole 20 ns trajectory. A

total of three trajectories with different random seeds was run for each sequence and the averaged B factors were calculated using the following equation: $B = (8\pi^2/3) \times (\text{RMSF})^2$, where the RMSF units are Å.

References

1. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278:82-87.
2. Pantazes RJ, Grisewood MJ, Maranas CD (2011) Recent advances in computational protein design. *Curr Opin Struct Biol* 21:467-472.
3. Röthlisberger D, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190-195.
4. Jiang L, *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387-1391.
5. Fleishman SJ, *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332:816-821.
6. King NP, *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336:1171-1174.
7. Grigoryan G, *et al.* (2011) Computational design of virus-like protein assemblies on carbon nanotube surfaces. *Science* 332:1071-1076.
8. Mandell DJ, Coutsiaris EA, Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* 6:551-552.

9. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using β -strand assembly. *Proc Natl Acad Sci USA* 108:20562-20567.
10. Fleishman SJ, *et al.* (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J Mol Biol* 414:289-302.
11. Whitehead TA, Baker D, Fleishman SJ (2013) Computational design of novel protein binders and experimental affinity maturation. *Method Enzymol*, ed Amy EK (Academic Press), Vol Volume 523, pp 1-19.
12. Lim KH, Dyson HJ, Kelly JW, Wright PE (2013) Localized structural fluctuations promote amyloidogenic conformations in transthyretin. *J Mol Biol* 425:977-988.
13. Blomberg R, *et al.* (2013) Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* doi:10.1038/nature12623.
14. Malakauskas SM, Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* 5:470-475.
15. Havranek JJ, Harbury PB (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* 10:45-52.
16. Tinberg CE, *et al.* (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501:212-216.
17. Koder RL, *et al.* (2009) Design and engineering of an O₂ transport protein. *Nature* 458:305-309.

18. Allen BD, Nisthal A, Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc Natl Acad Sci USA* 107:19838-19843.
19. Kiss G, Röthlisberger D, Baker D, Houk KN (2010) Evaluation and ranking of enzyme designs. *Protein Sci* 19:1760-1773.
20. Privett HK, *et al.* (2012) Iterative approach to computational enzyme design. *Proc Natl Acad Sci USA* 109:3790-3795.
21. Liu Y, Eisenberg D (2002) 3D domain swapping: As domains continue to swap. *Protein Sci* 11:1285-1299.
22. O'Neill JW, Kim DE, Johnsen K, Baker D, Zhang KY (2001) Single-site mutations induce 3D domain swapping in the B1 domain of protein L from *Peptostreptococcus magnus*. *Structure* 9:1017-1027.
23. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D (2001) Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc Natl Acad Sci USA* 98:10687-10691.
24. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO (1994) Structural studies of the engrailed homeodomain. *Protein Sci* 3:1779-1787.
25. Huang P-S, Love JJ, Mayo SL (2005) Adaptation of a fast Fourier transform-based docking algorithm for protein design. *J Comput Chem* 26:1222-1232.
26. Rekas A, Alattia JR, Nagai T, Miyawaki A, Ikura M (2002) Crystal structure of Venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity. *J Biol Chem* 277:50573-50578.

27. Choi EJ, Mayo SL (2006) Generation and analysis of proline mutants in protein G. *Protein Eng Des Sel* 19:285-289.
28. O'Neill JW, Kim DE, Baker D, Zhang KYJ (2001) Structures of the B1 domain of protein L from *Peptostreptococcus magnus* with a tyrosine to tryptophan substitution. *Acta Crystallogr D* 57:480-487.
29. Liang S, *et al.* (2008) Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. *Biochemistry* 48:399-414.
30. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435-447.
31. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. *Science* 334:517-520.
32. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography* 60:2126-2132.
33. Adams PD, *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D-Biological Crystallography* 66:213-221.
34. Shah PS, *et al.* (2007) Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol* 372:1-6.

Table 2-1. Sequences of wild-type ENH, ENH_DsD, E23P, and E24P



ENH (WT)	TAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI
ENH_DsD	-E--E--KKA-DLA-YFD-R---EW-RY--QR-----E--ER--RR-EQQ-
E23P	-E--E--KKA-DLA-YFD-R--PEW-RY--QR-----E--ER--RR-EQQ-
E24P	-E--E--KKA-DLA-YFD-R---PW-RY--QR-----E--ER--RR-EQQ-

The three “coils” at the top show the location of the three helices in the wild-type (WT) fold.

Table 2-2. Summary of X-ray crystallography, dimerization assays, CPD calculations, and MD-derived hinge B factors for each of the proteins tested.

	Domain-	K_d^*	CPD energy [†]	B-factor [‡]
ENH (WT)	No	n/a	-127.0	21.1
ENH_DsD	Yes	40nM	-125.2	70.9
E23P	No	n/a	-123.7	32.7
E24P	n/a	n/a	-120.6	106.6
1HZ5 (WT)	No	n/a	-183.1	78.4
G55A	Yes	30uM	-166.2	58.2
A52V/D53P/G55A	Yes	700pm	249.1	179.6

*Dissociation constant for the domain-swapping dimer.

[†]CPD energy was calculated in the context of the monomer structure.

[‡]The maximal B-factor on the hinge.

Table 2-3. Data collection and refinement statistics for the crystal structure of ENH_DsD-YFP (PDB: 4NDJ)

Statistics	Value
Data collection	
Space group	P4222
Cell dimensions	
<i>a, b, c</i> (Å)	104.4, 104.4, 80.2
<i>α, β, γ</i> (°)	90.0, 90.0, 90.0
Resolution (Å)	30–1.85
R _{merge}	0.052
I/σ	20.2
Completeness, %	99.2
Multiplicity	7.3
Refinement	
Resolution (Å)	30–1.85
Number of reflections	39983
R _{work} /R _{free} (%)	17/20
Number of molecules in asymmetric unit	1
Number of atoms	2627
Protein	2275
Water	333
Ligands	19
B factors (Å ²)	31.2
Protein	29.7
Water	42.0
Ligands	21.3
R.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.16
Ramachandron map analysis	
Most favored regions (%)	98.9%
Additional allowed regions (%)	1.1%
Disallowed regions	0%

Table 2-4. Data collection and refinement statistics for the crystal structure of E23P-YFP (PDB: 4NDK)

Statistics	Value
Data collection	
Space group	P222 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	75.9, 192.7, 107.6
α , β , γ (°)	90.0, 90.0, 90.0
Resolution (Å)	34–2.3
R _{merge}	0.064
I/ σ	8.3
Completeness, %	99.4
Multiplicity	4.0
Refinement	
Resolution (Å)	34–2.3
Number of reflections	35208
R _{work} /R _{free} (%)	19/24
Number of molecules in asymmetric unit	2
Number of atoms	4848
Protein	4416
Water	394
Ligands	38
B factors (Å ²)	35.0
Protein	34.9
Water	36.6
Ligands	24.5
R.m.s. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.15
Ramachandron map analysis	
Most favored regions (%)	96.4%
Additional allowed regions (%)	3.4%
Disallowed regions	0.2%

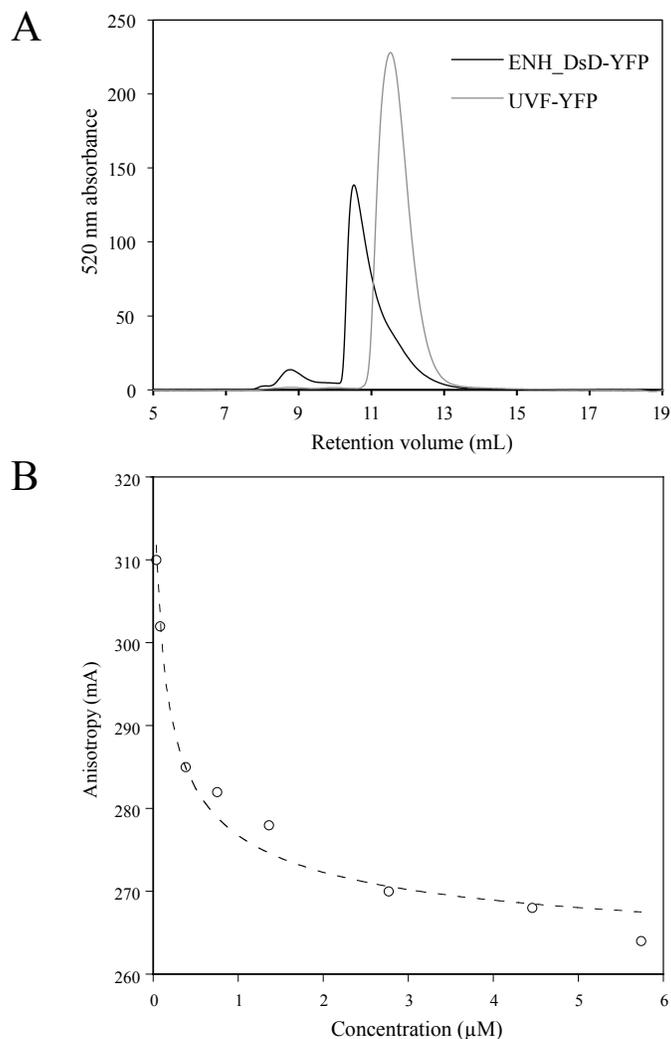


Fig. 2-1. Size exclusion and polarization fluorescence assays indicate that ENH_DsD-YFP is a high-affinity dimer. (A) Size-exclusion chromatography for ENH_DsD-YFP (black) and a monomeric control, UVF-YFP (gray). Absorbance at A515 (absorbance peak for YFP) was tracked for protein elution. UVF is a computationally designed 39-fold mutant of ENH whose NMR solution structure matches the wild-type (monomeric) fold (34). The retention volumes for UVF-YFP and ENH_DsD-YFP are consistent with standards for a monomer and a dimer, respectively. (B) Polarization fluorescence of ENH_DsD-YFP. The unit of anisotropy (mA) is a thousandth of anisotropy (A). A K_d of ~40 nM was calculated based on a simple monomer-dimer equilibrium model.

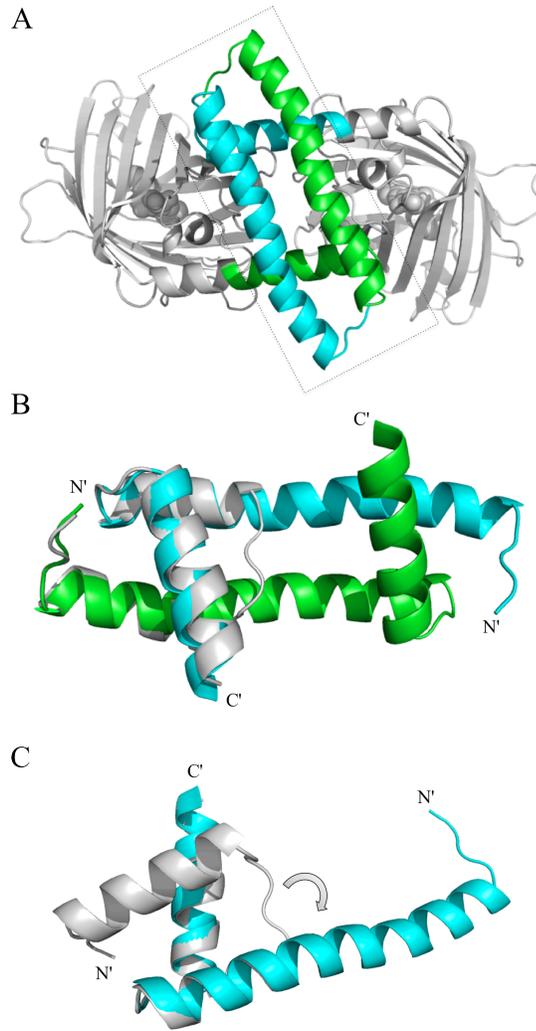


Fig. 2-2. X-ray crystallography reveals that ENH_DsD-YFP is a domain-swapped dimer. (A) Crystal structure of ENH_DsD-YFP resolved to 1.85 Å. The two chains in ENH_DsD are shown in green and cyan, respectively, and the YFP sequences are shown in gray. (B) Zoom-in of the dashed frame in A with the wild-type ENH structure (PDB ID: 1ENH) (gray) superimposed. (C) The hinge-loop between the first and second helices in the wild-type structure (gray) has flipped over by $\sim 90^\circ$ (arrow) and coiled up to form the domain-swapped conformation seen in ENH_DsD-YFP (cyan). This change leads to a $\sim 180^\circ$ rearrangement of helix 2 so that helix 1, the hinge, and helix 2 now form a single long helix.

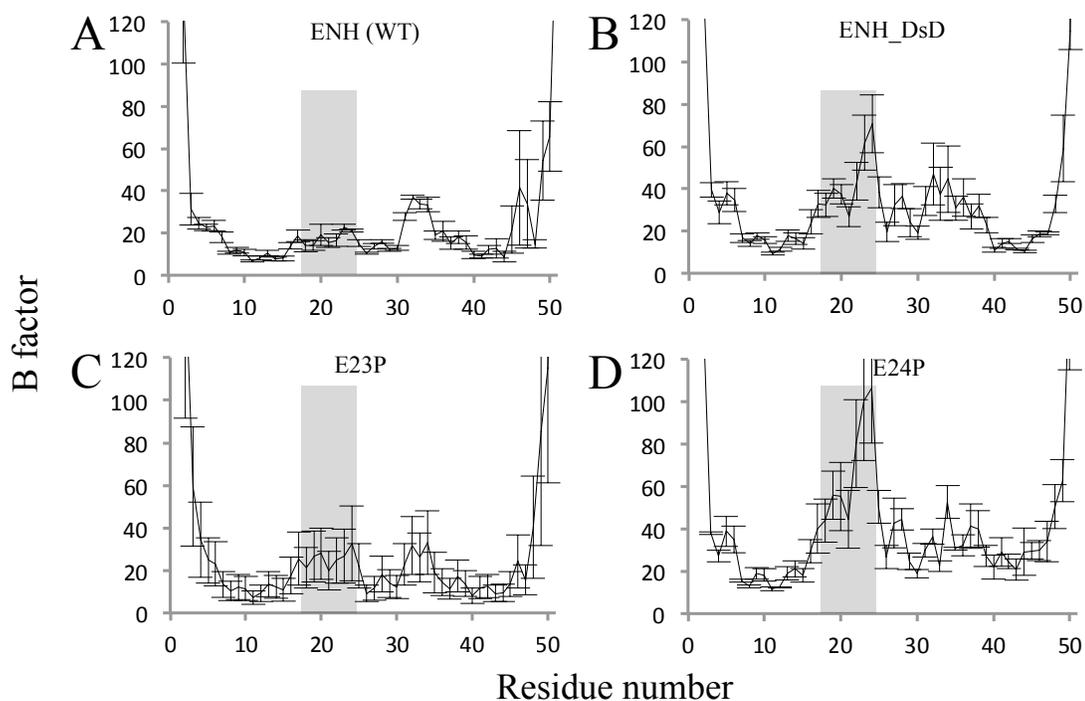


Fig. 2-3. B factor analyses from MD simulations for wild-type ENH (*A*), ENH_DsD (*B*), E23P (*C*), and E24P (*D*). 20 ns MD simulations were run on wild-type ENH and on each of the variant sequences threaded onto the wild-type backbone structure, the RMSF of the C_{α} atoms was analyzed using GROMACS for each of three trajectories, and the averaged B factors were calculated. Error bars: SD for three independent trajectories. Hinge residues (between helices 1 and 2 in wild-type ENH) are indicated with a gray bar in each panel.

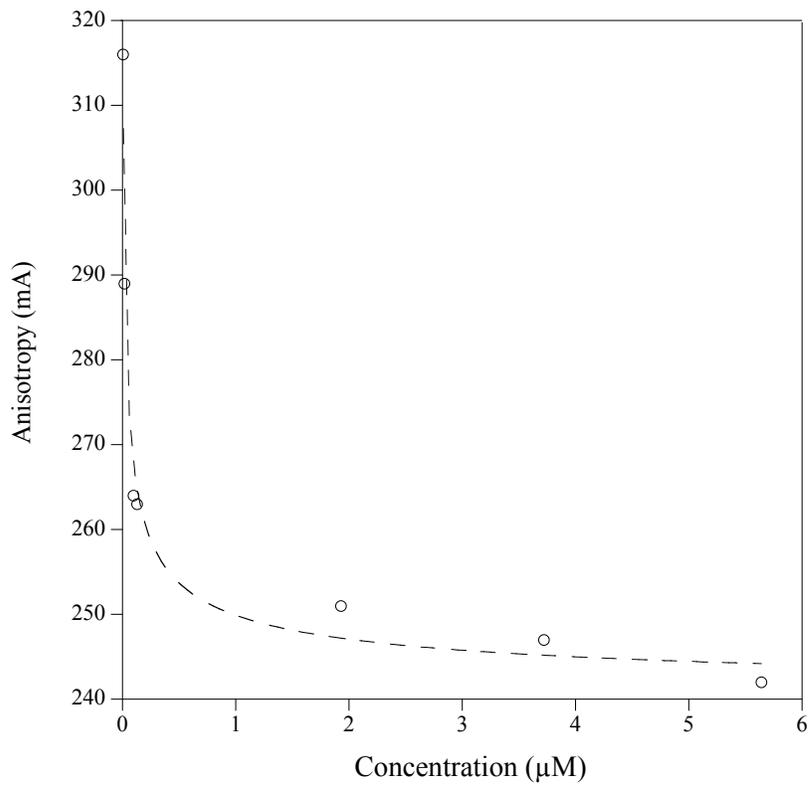


Fig. 2-4. Polarization fluorescence indicates that E23P-YFP is a high-affinity dimer. Polarization fluorescence of E23P-YFP was measured with a series of diluted samples in buffer containing 100 mM NaCl and 20 mM TrisHCl at pH 8.0. The unit of anisotropy (mA) is a thousandth of anisotropy (A). The G-factor was determined for each data point. A K_d of ~10 nM was calculated based on a simple monomer-dimer equilibrium model.

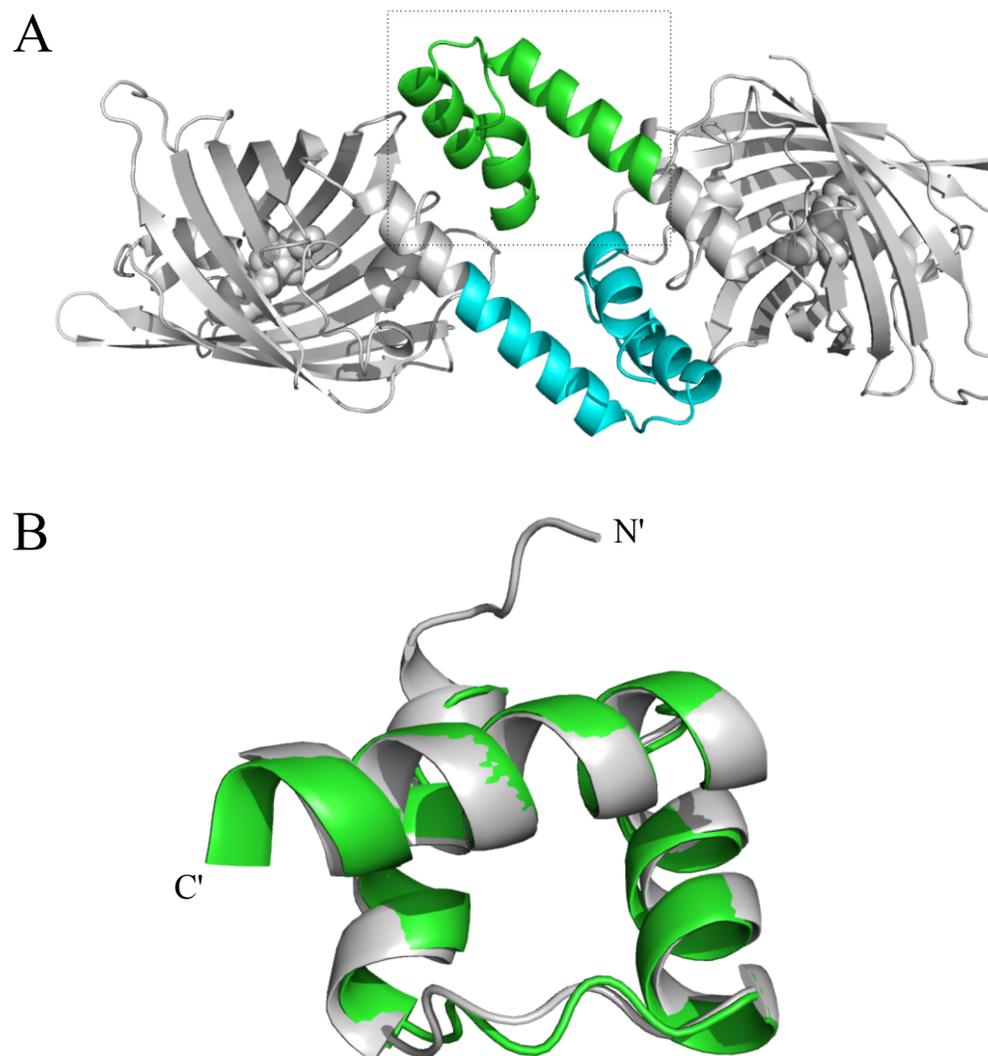


Fig. 2-5. X-ray crystallography of E23P-YFP shows that a single proline mutation in the hinge recovers the wild-type fold. (A) Crystal structure of E23P-YFP resolved to 2.3 Å. The E23P sequence is shown in green and cyan (for each of the two chains, respectively) and the YFP sequences are shown in gray. (B) Zoom-in of the dashed frame in A with the wild-type ENH structure (gray) superimposed. E23P exhibits the wild-type fold.

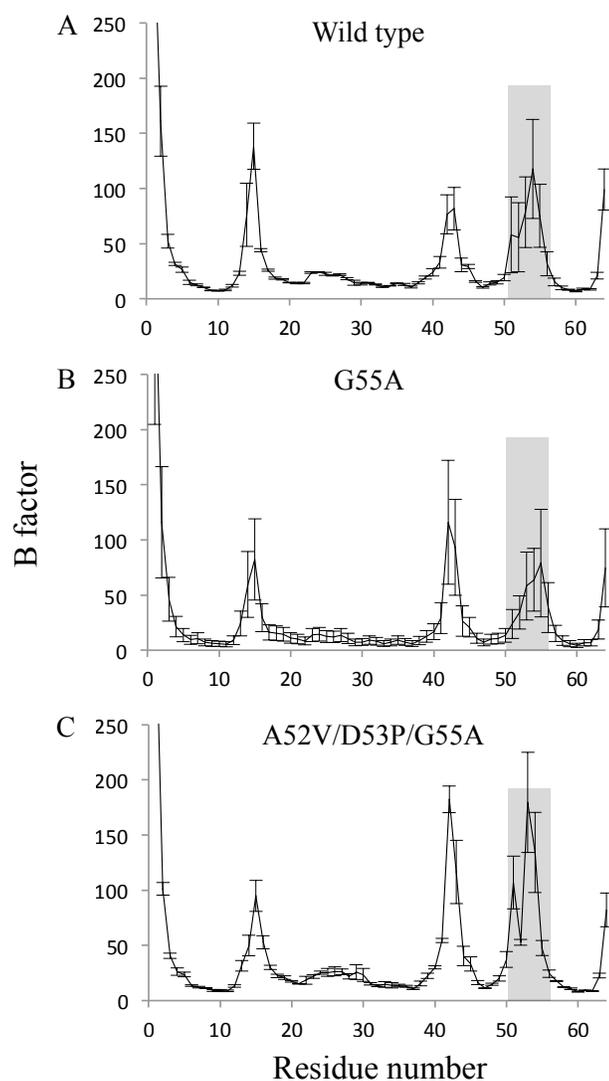


Fig. 2-6. B factor analyses from MD simulations for the B1 domain of protein L (wild-type) (*A*), G55A (*B*), and the triple mutant A52V/D53P/G55A (*C*). For G55A and the triple mutant, sequences were first threaded onto the wild-type monomeric structure. As described above, 20 ns MD simulations were run, the RMSF of the C_{α} atoms was analyzed using GROMACS for each of three trajectories, and the averaged B factors were calculated (see Methods for details). Error bars: SD for three independent trajectories. Hinge residues are indicated with a gray bar in each panel. The backbone dynamics of the hinge varies for the different proteins and reflects their domain-swapping tendencies.

Chapter 3

Computational Design and Experimental Verification of a Symmetric Homodimer

Abstract

Homodimers are by far the most common type of protein assembly in nature and have distinct features compared to heterodimers and higher-order oligomers. Understanding homodimer interactions at the atomic level is critical both for accurate modeling and for elucidating their biological mechanisms of action. Computational design of novel protein-protein interfaces can serve as a bottom-up method to further our understanding of protein interactions. Previous studies have demonstrated that the *de novo* design of homodimers can be achieved to atomic-level accuracy by β -strand assembly or metal mediation. Here, we report a novel homodimer with C2 symmetry that is designed via helical interactions. The structure obtained by solution nuclear magnetic resonance shows that the homodimer exhibits parallel helical packing similar to the designed model. Because designing for improved functionality often results in decreased thermostability, a stability design step was introduced. In addition to our standard docking and design procedure, to compensate for the poor thermostability of the scaffold. This two-step design approach is essential when a thermophilic protein is not available or desirable as the scaffold for design.

Introduction

Protein-protein interactions play a central role in nearly all biological processes, including cell signaling, immune responses, regulation of transcription and translation, and cell-cell adhesion. Improving our understanding of protein-protein interactions is therefore crucial to advancements in both basic and applied research in the pharmaceutical, chemical, and biotechnology industries. The influx of structures deposited into the Protein Data Bank (PDB) during the last two decades has helped researchers identify unique features of protein-protein interactions (1-3). Specific interfacial residues that contribute to most of the binding energy (“hot spots”), networks of hydrogen bonds, and shape complementarity have all been identified as important. These features have therefore been incorporated into many protein docking and protein design algorithms (4, 5). Protein docking algorithms have successfully been used to screen millions of docking positions and identify the correct (near-native) structure (6). Computational design tools have also exploited our knowledge of protein-protein interactions to successfully design enhanced affinity or altered specificity (7, 8), to graft binding motifs onto a desired scaffold (9, 10), and to create novel interfaces (11-16). Several studies have shown the atomic-level accuracy of *de novo* designed protein dimers. However, the success rate of novel interface designs is rather low (17) and remains one of the most challenging areas in the field (18).

Homodimers are by far the most common type of protein assembly and are well represented in the PDB. Compared to heterodimers, homodimers have larger surface area, fewer hydrogen bonds, higher hydrophobicity, and typically, C2 symmetry (19). Although homodimers are abundant in nature, there are only a few examples of the computational design of symmetric homodimers. Stranges et al. showed that solvent-exposed β -strands can be used as anchors to design a symmetric homodimer that associates via β -strand pairing (16). Der et al. incorporated metal binding sites to drive homodimerization and achieve high affinity and orientation specificity (13). Interestingly, in their study, the helices on each side of the metal-mediated homodimer interface aligned nearly orthogonally, unlike the parallel or anti-parallel alignments of helices typically found

in nature. Both parallel and anti-parallel coiled-coil like dimers have been designed using short peptides (20, 21). However, there have been no structurally verified homodimers designed via novel helical interfaces between proteins. Helical interactions, often in the form of coiled-coils, occur twice as frequently at homodimer interfaces (22.4%) compared to heterodimer interfaces (10.9%), but strand-strand interactions are seen at about the same frequency (8.8% and 8.4%, respectively) (22). We were therefore very interested to see whether we could computationally design a helical interface on a monomeric globular protein that facilitates homodimerization with C2 symmetry.

Designing a homodimer by helical interactions presents many challenges. First, unlike strand-strand interactions, where association occurs via specific backbone hydrogen bonds, the helical interface does not provide chemically specific anchors for protein-protein interactions. Although there are some empirical rules for archetypal coiled-coil oligomerization (23), a general sequence-structure relationship that could be applied to any arbitrary scaffold has not yet been found. Moreover, high-throughput methods for identifying protein-protein interactions (e.g., yeast display) cannot be used for homodimer screening, as heterodimers will dominate a library. Furthermore, as shown by Keating and co-workers, predicting parallel or anti-parallel helix-helix homodimers using computational modeling is not trivial (24). The similarity between parallel and anti-parallel helix-helix structures and the high hydrophobicity of homodimers make it difficult to distinguish between the different conformational states, particularly if they are strongly competing with each other and only one of the states is explicitly designed. For example, Karanicolas et al. computationally designed a novel protein-protein interface with tightly packed hydrophobic residues (25). The crystal structure, however, revealed that the orientation of one of the partners was rotated almost 180° relative to its position in the design model. These results underscore the difficulty of excluding unwanted competing states in the design of protein-protein interactions.

Here, we design a C2-symmetry homodimer from a helical monomeric protein, *Drosophila melanogaster* engrailed homeodomain (ENH). This small helix-turn-helix protein binds a specific sequence of double-stranded DNA (dsDNA) (26), and has been used as a model for many

theoretical and computational studies (27-29). Computational protein design (CPD) projects often begin with a highly thermostable scaffold, because designs for improved function (e.g., catalytic activity, ligand-protein binding affinity) can decrease protein stability significantly (30). Poor stability in turn often results in aggregation and can be problematic for recombinant expression and/or experimental characterization. Wild-type ENH has a low melting temperature (T_m) of 49°C (31). Indeed, we found that interface design of wild-type ENH led to a protein (ENH-c2a) that expresses in inclusion bodies even at reduced temperatures. We then incorporated mutations from Marshall et al. (31) that achieved improved thermostability (an ENH variant called NC3-NCap with $T_m = 89^\circ\text{C}$) for later designs. We applied a symmetrical docking program based on a fast Fourier transform algorithm (32) and designed the interface between four helices (two from each molecule) so they would associate as a four-helix bundle. The final design (ENH-c2b) was experimentally characterized as a monodisperse homodimer with a K_d of ~130 nM. The solution nuclear magnetic resonance (NMR) structure reveals that the helical interface exhibits parallel packing, as designed in our model.

Results

Scaffold Selection. Due to our incomplete understanding of structure-function relationships and limitations in our ability to accurately model proteins, the success rate for computational protein design is rather low. This is particularly true for the *de novo* design of functional proteins such as enzymes with novel catalytic activity or proteins designed to interact with a specific target (33). Many *de novo* design studies therefore use a comprehensive approach that virtually screens all favorable scaffolds in the PDB, often resulting in tens to hundreds of candidate scaffolds (11, 16, 34, 35). These candidates are then screened experimentally (e.g., for target binding affinity, catalytic activity). This approach has the advantage of providing a large amount of data that can be used to test design protocols (18, 36). However, it usually relies on a general design strategy, which may not be optimal for any particular scaffold. Alternatively, using a single scaffold allows the design to

take specific features of the scaffold into account. For example, Privett et al. used an iterative approach on a single scaffold that resulted in the *de novo* design of the most efficient Kemp eliminase to date (30). Scaffold properties such as the hydrophobicity of the active site pocket and substrate packing were taken into account in second and third generation designs, yielding improved enzyme variants. Choosing a particular scaffold may also be desirable if one is interested in certain native properties of the protein that are crucial to one's ultimate design goals. We chose ENH, for instance, for its DNA-binding properties, which would facilitate our final goal of creating a protein that can loop DNA or form a protein-DNA nanoparticle. The co-crystal structure of ENH and its target DNA shows that this scaffold can allow for homodimer design while retaining its DNA-binding capability. The third helix (helix-3) is the major DNA binding domain. The first and second helices (helix-1 and helix-2) are not in contact with the DNA (26), and expose a large flat surface that is potentially appropriate for designing a homodimer interface via helical interactions. ENH was therefore chosen as our single starting scaffold for docking and homodimer design.

Design Protocol. Fig. 3-1 shows the steps we used to design and characterize a C2-symmetrical homodimer. After selecting the scaffold protein, the surface sidechains were pruned to C β and the atomic radii were parameterized based on known C2 symmetry homodimers. We then applied a symmetrical docking program that we had developed based on a fast Fourier transform (FFT) algorithm (32). FFT docking allowed us to efficiently search all six translational and rotational docking positions. About 10^{10} docking positions were screened and ranked by shape complementarity. The top 200 candidates were clustered into 11 groups according to the root-mean-squared deviations (RMSDs) of the structures. Finally, these clusters were visually inspected and one model was chosen for homodimer design. In this homodimer model, the two helix-1's (one from each subunit) form parallel helix-helix packing and are separated by 10 Å, similar to the 9.8 Å separation found in naturally occurring coiled-coil dimers (37). Note that our design process did not explicitly use the heptad repeat rules that govern coiled-coil interactions (20); the designed interface

is determined by the physical-chemical interactions specified by the force field. The two helix-2's in the dimer are on opposite sides of the two helices 1 (one on each side). Together, helix-1 and helix-2 from the two subunits form a four-helix bundle structure (see bronze model, Fig. 3-1).

Next, symmetrical sequence optimizations were applied to the interfacial residues of the homodimer model. To mimic natural homodimers, which have relatively high hydrophobicity at the interface (~65%) (19), the balance between polar and non-polar amino acids was carefully monitored during sequence optimization. In the force field used for our optimizations, the energy term that is predominantly responsible for hydrophobicity is the solvation energy (38):

$$E_{as} = \sigma_{np}A_{np,b} + \sigma_{np}A_{np,e} + \sigma_p A_{p,b} ,$$

where the atomic solvation energy (E_{as}) consists of a benefit term for burial of nonpolar amino acids ($\sigma_{np}A_{np,b}$) and penalty terms for non-polar amino acid exposure ($\sigma_{np}A_{np,e}$) and polar amino acid burial ($\sigma_p A_{p,b}$). Each of these terms contains a solvation energy factor (σ) and a surface area component (A). The value of σ_p specifies the magnitude of the penalty for burying polar atoms. By adjusting this factor, we fine-tuned the hydrophobicity of the designed interface to be as close to natural homodimers as possible. The final value of σ_p for our model was determined to be 0.04, raised by 0.02 from its standard value.

Homodimer design was first done using wild-type ENH as our starting scaffold. Unfortunately, the resulting protein (ENH-c2a) could only be expressed in inclusion bodies (Fig. 3-2) so we repeated the interface design using a thermostabilized variant of ENH that had been generated previously (NC3-NCap) (31) (the NC3-NCap sequence was threaded onto the docked homodimer model). We expected that this more stable protein would result in designed sequences with improved soluble expression. In order to examine the top designed sequences in a high-throughput manner, a computational library consisting of 128 variants of NC3-NCap was generated (Table 3-1), fused to YFP, and screened using a homo Förster resonance energy transfer (FRET) assay (Fig. 3-3A). Two representative high-affinity variants resulting from the screen were

characterized via size exclusion chromatography and sedimentation velocity experiments; one proved to be a dimer, and the other was a tetramer (Fig. 3-3B and C). In Chapter 2, we showed that the dimer is domain-swapped (ENH_DsD), and that we could revert it to the wild-type monomeric fold by substituting residue 23 with a proline, thereby decreasing hinge flexibility. The resulting sequence (ENH-c2b) expressed well in the soluble fraction. The sequences of the two homodimer designs (ENH-c2a and ENH-c2b) and their starting scaffolds (ENH and NC3-NCap, respectively) are listed in Table 3-2.

Biophysical Characterization of ENH-c2b. The designed proteins were characterized for soluble expression, secondary structure, thermostability, and oligomeric state. SDS-PAGE gels of purified ENH, ENH-c2a, and ENH-c2b are shown in Fig. 3-2. Although ENH could be expressed in the soluble fraction at 16°C, ENH-c2a showed no soluble expression under the same conditions. In contrast, ENH-c2b expressed well at 37°C, with yields of over 5 mg per liter culture. Circular dichroism (CD) spectroscopy revealed that ENH-c2b is a helical protein with a perfectly reversible denaturing curve that has a T_m of ~62°C (Fig. 3-4A and B). The designed interface reduced the T_m by ~26°C from 88°C, which is the T_m of the starting scaffold (NC3-NCap). The much lower thermostability of the starting scaffold used to generate ENH-c2a (ENH, T_m = 49°C) might explain why this protein failed to be solubly expressed.

Size-exclusion chromatography of ENH-c2b showed fast reversible equilibrium between the monomeric and dimeric forms (Fig. 3-5A). The elution peak occurred earlier with high concentration loading (14 mL) than with low loading (16 mL). Using standard elution profiles, we assigned the elution peaks at 14 and 16 mL as the dimeric (~13 kD) and monomeric states (~6.5 kD), respectively. Sedimentation velocity experiments showed that ENH-c2b is a monodisperse dimer at 5 μ M concentration (Fig. 3-5B), which implies that the K_d is in the sub- μ M range or lower. The K_d was determined using a tryptophan fluorescence-based homo-FRET assay, and was found to

be 129 ± 64 nM (Fig. 3-5C), which is similar to the K_d values reported for other *de novo* protein interface designs (pre-affinity maturation) (11, 13, 16).

Structural Determination of ENH-c2b. Extensive trials failed to generate high-quality crystals of ENH-c2b for X-ray diffraction. However, we did obtain a crystal structure of a variant that included an extra 21 residues at the N-terminus, but it turned out to be monomeric, as determined by PISA (39) (see Fig. 3-6). Next, we attempted to solve the structure of ENH-c2b using solution NMR. Heteronuclear single quantum coherence (HSQC) of freshly-prepared ENH-c2b showed a well-folded protein with sharp peaks, but the peaks broadened over time. Adding a glycine and an 8-residue Strep • Tag II at the C-terminus of ENH-c2b (ENH-c2b-Strep) greatly enhanced its long-term stability. We were able to unambiguously assign chemical shifts to almost all of the backbone nuclei. However, peaks were missing for residues 21-23 (FYF) at the end of helix-1. This is consistent with the fact that the chemical shifts of aromatic residues are highly sensitive to their sidechain conformations, and can be easily broadened if multiple conformations exchange quickly.

We determined the structure of the ENH-c2b-Strep homodimer using ψ/ϕ angle, hydrogen-bond, Nuclear Overhauser Effect (NOE), and C2-symmetry restraints (Fig. 3-7, green structure). Final coordinates were deposited in the PDB with code 2MG4. Each monomeric subunit is superimposable with wild-type ENH except for the long loop between helices 1 and 2, residues 21-23 (at the end of helix-1), and the N termini (Fig. 3-7A and B). Restraints are lacking for these regions, so they appear disordered in the NMR ensemble. Note that the interface helices (helix-1 and helix-2) align almost perfectly with the design model (Fig. 3-7A). However, the orientation of helix-3 deviates slightly from that of the model (Fig. 3-7B). Compared to ENH, the rigid fragments of ENH-c2b-Strep (helices 1, 2, and 3, and the loop between helix-2 and helix-3) are very similar (backbone RMSD = 1.35 Å). Note that the sequence identity between ENH-c2b and ENH is only 47%.

The solution structure of the ENH-c2b-Strep dimer shows parallel helix-helix packing between helix-1 of each subunit, as in our homodimer model. Compared to our model, the backbone RMSD of the rigid fragments is 2.23 Å (Fig. 3-7C and D). The axial orientations of the four helices (helices 1 and 2 from each of subunits) are nearly identical with those in the model (Fig. 3-7C). The interface area is 2189 Å², which falls in the range of natural single-patch homodimers (2740 ± 1240 Å²) (19). Nonpolar residues constitute 62% of the interface, which is very close to the average value of 65% for natural homodimers (19). A NOESY experiment designed to retain only inter-molecular interactions revealed several nonpolar interfacial residues that are likely to be important for dimerization, including Ala16, Leu19, Ala20, and Leu39 (Fig. 3-8). Structural alignment of only one of the subunits emphasizes that there are some differences from the model (Fig. 3-7E and F). These variations are expected given that the main driving force for dimerization in our design was hydrophobic interactions, which are less specific than other interactions.

Discussion

Stranges et al. recently reviewed the computational design of novel protein-protein interfaces and pointed out the challenges that this burgeoning field faces (17). Of 147 protein-protein interaction designs, only four were confirmed successful by X-ray crystallography (i.e., the solved structure matched the design model). All the successful designs shared some common features—they exhibited fewer polar atoms (< 40%) and fewer buried hydrogen bonds at the designed interface than those seen in the failed designs. This reduced number of buried hydrogen bonds in the successful designs is in contrast to what is typically observed in natural dimers. Our homodimer design showed similar results (62% nonpolar atoms and no buried H-bonds at the interface). This higher interfacial hydrophobicity is expected given that three of the five successful cases (including ours) were designs for homodimers (13, 16), which naturally exhibit high nonpolar content at the interface. In our case, we purposely biased the interface to replicate this naturally high nonpolar content by adjusting the solvation energy term. The other two successful cases

(heterodimer designs) also exhibited high interfacial hydrophobicity, probably because the targeted surface, the stem region of influenza hemagglutinin (HA), is a highly hydrophobic patch (11, 15). Hydrophobic surfaces alone often confer little specificity. For example, the designed protein could bind to the target protein in the wrong orientation (25) or bind to itself to form undesired oligomers (11). Nonetheless, the design of protein-protein interfaces via hydrophobic interactions has led to successful results, whereas the successful design of polar interfaces still eludes us. Thus far, there are few reports of protein-protein interactions designed via hydrophilic interfaces. Very recently, Procko et al. designed a protein inhibitor that binds to a hydrophilic patch on lysozyme with high affinity; however, this complex has yet to be structurally validated (40).

Another interesting feature of all five successful designs is that the designed interfaces mainly involve secondary structures. Both of the successful HA heterodimer designs described above have hot spots on their helices that bind to helical structures on the target HA stem. Of the three successful homodimer designs to date, one is between two helices (this work), one involves metal-protein interactions on the helices (13), and one uses two exposed β -strands to form the homodimer (16). Loops can also be exploited in protein-protein interactions, as demonstrated by the widespread use of loops in antibody-antigen interactions. Many computational loop designs have been attempted, but thus far, none have resulted in dimerization (11). In a community-wide assessment of protein-protein designs, Haliloglu and coworkers found that many of the failed designs contain more loops and turns than the successful ones, and the higher flexibility of these structures makes adopting a particular designed conformation difficult (18). Non-interfacial loop designs have occasionally proved successful (41-43). However, the design of interfaces involving loops appears to be more challenging, as the recognition-induced conformational changes that loops undergo upon association with another protein are still poorly understood and poorly modeled. Given these results, we suggest that future computational protein-protein designs include an *in silico* screening step that eliminates docked dimer models in which loops make up a significant portion of the interface.

One difference between our design and the four successful designs reported previously is that the accuracy of our design is somewhat lower than the others (RMSD for our design = 2.23 Å, whereas that for the other designs ranged from 1.0 to 1.8 Å). The success of the four more accurate designs may be due to their incorporation of specific hot-spot residues or anchoring interactions that steered the formation of a high-affinity dimer. For example, the two HA heterodimer designs used pre-defined hot-spot residues to match the specified locations on the target patch (11), and the homodimer design employed β -strand hydrogen bonds (16) or metal chelators to anchor the homodimerization (13). All of these designs exploited very specific pair-wise interactions to facilitate and guide complex formation. In addition, they also incorporated relaxation into the design process to fine-tune docking positions. Although our design used a fixed backbone for sequence optimization and did not specify hot spots to orient docking, we still obtained a homodimer that matched our model relatively well.

This work presents a CPD approach in which the stability and the functionality of the protein are taken into account in a stepwise fashion. Conventionally, CPD uses protein scaffolds with known structures for sequence designs. Here, we initially used wild-type ENH for homodimer design, but the resultant protein (ENH-c2a) could not be solubly expressed. We then decided to use a thermostabilized variant of ENH (NC3-NCap) as our starting scaffold. Although NC3-NCap showed satisfactory properties, including improved thermostability, and secondary structure similar to wild-type ENH, its three-dimensional structure has not yet been confirmed by X-ray crystallography or NMR. Nevertheless, we modeled NC3-NCap using the wild-type ENH fold and proceeded with the interface design. Encouragingly, our resultant protein, ENH-c2b, proved to be homodimeric, as designed, and was shown by NMR to share the same fold as its parent ENH. Thus, in this case, stability design apparently facilitated the successful functional design of a homodimer. This “two-step” design approach is illustrated in Fig. 3-9. During the last decade, CPD has proved to be very powerful for designing proteins with improved stability (44, 45). Recently, the field has moved to include the design of functionality, such as new or improved catalytic activity, binding

affinity, ligand or substrate specificity, etc. Functionality design can, however, destabilize a protein significantly (30). In the worst-case scenario, the designed protein variants cannot be solubly expressed. For example, Fleishman et al. designed 88 proteins to bind to HA and found that 50% of them could not be solubly expressed in *Escherichia coli* (5). In our work here with ENH, we showed that an apparent trade-off between stability and functionality could be solved by adding a stability design step to the design process, and that structural validation of the thermostabilized variant was not required to generate a successful C2-symmetry homodimer.

Conclusions

This work represents the first *de novo* design of a C2-symmetry homodimer via helical interactions. The successful design (ENH-c2b) is a monodisperse dimer with a K_d of 130 nM. The solution NMR structure is generally consistent with our design model, with the protein-protein interface forming a four-helix bundle. The homodimer design was achieved using a two-step computational approach in which the wild-type protein (ENH) was first stabilized and subsequently homodimerized. The final design (ENH-c2b) is 13°C more stable than ENH, and its sequence identity is 47%. This value is notably lower than for other interface designs (80-90%) and can be accounted for by the large number of mutations (34) that resulted from stabilization. Our successful homodimer design illustrates that the judicious use of computational tools can be employed to generate a stable protein with the desired functionality.

Methods

Protein Docking and Computational Design. The ENH crystal structure (PDB code: 1ENH) was used as the scaffold for homodimerization, with side-chain atoms beyond C_β deleted and atomic radii of the remaining atoms adjusted as follows: N: 1.4 Å, O: 1.3 Å, C': 1.75 Å, C_α : 2.35 Å, and C_β : 2.15 Å. A symmetrical docking program based on a fast Fourier transform (FFT) algorithm was applied (32). Arrays used for docking calculations were $64 \times 128 \times 128$ for each of the x, y, and z dimensions, with each element corresponding to 1 cubic Å. Each round of searching consisted of

extensive translational dockings followed by 1° increments about the y and z axes. Using shape complementarity as the criterion, the top 20 conformations for each rotational position were identified, combined into a set containing all the top 20s, then ranked. The top 200 of these were clustered into 11 groups based on structural similarity (RMSDs), the clusters were visually inspected, and one high-scoring model was selected for computational designs. The ORBIT computational protein design (CPD) software was used for stability designs for both ENH and NC3-NCap. Initial interface designs were also done using ORBIT, and subsequent designs and analyses were done using our improved CPD programs, PHOENIX and Triad. Sequence optimization was performed using an improved version of FASTER (refs) and a rotamer library based on the backbone-dependent library of Dunbrack and Karplus (ref).

Construct Preparation, Protein Expression and Purification. Oligonucleotides (Integrated DNA Technologies) containing ~20 bp overlapping segments were assembled via a modified Stemmer polymerase chain reaction (PCR) method (ref) using KOD Hot Start Polymerase (Novagen) to generate genes for ENH, ENH-c2a, and ENH-c2b. For SDS-PAGE analysis, a His₆ or Strep • Tag II was added to generate His₆-ENH, ENH-c2a-Strep, and ENH-c2b-Strep constructs. For all other biophysical characterizations, intact ENH-c2b was used. X-ray crystallography was performed on an ENH-c2b variant containing 21 additional residues at the N terminus (MGSSHHHHHSSGLVPRGSHM). The construct for solution NMR was ENH-c2b, with an extra C-terminal glycine followed by a Strep • Tag II. All proteins were expressed using BL21 DE3 cells transformed by pET plasmids with 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG) in standard Luria Broth (LB) at 16°C (His₆-ENH and ENH-c2a-Strep) or 37°C (all other proteins). The ¹³C/¹⁵N labeled ENH-c2b-Strep for NMR experiments was prepared by growing BL21 DE3 cells in 1 L LB until OD₆₀₀ reached ~0.6 and transferring the cells to 250 mL M9 medium with ¹³C glucose and ¹⁵N ammonium chloride. Purification of ENH-c2b was accomplished by fusing it to His₆-ubiquitin, running the construct on a Ni²⁺-NTA column (Qiagen), then cleaving His₆-ubiquitin off using UCH-

L3 protease (37°C overnight). Strep-Tactin Sepharose (IBA) and Superdex 75 (Amersham Pharmacia) columns were used for Strep-tag affinity chromatography and size-exclusion chromatography, respectively.

Circular Dichroism Spectroscopy. CD studies were performed on an Aviv 62A DS spectropolarimeter equipped with a thermoelectric temperature controller. Samples were prepared in 100 mM sodium chloride and 20 mM sodium phosphate buffer at pH 7.5. Wavelength scans and temperature denaturations were carried out in cuvettes with a 0.1 cm pathlength at a protein concentration of ~10 μ M. Three wavelength scans were performed at 25°C for each sample and averaged. The thermal denaturation curve was collected at 222 nm from 0°C to 99°C, sampling every 1°C separated by 2 min equilibration times (signal averaging time was 1 sec). The refolding curve was collected after the thermal denaturation experiment using the same sample.

Analytical Ultracentrifugation. ENH-c2b was analyzed on an XL-1 analytical ultracentrifuge equipped with an AnTi60 rotor (Beckman Coulter). Two-channel epon-filled centerpieces were used for the sedimentation velocity experiment. Cells were torqued to 130 lb-inch and run at 60,000 rpm. Data were acquired at 230 nm and 20°C in continuous mode. Data were first fit to the c(s) model (continuous distribution of sedimentation coefficient) and then converted to the c(M) model (continuous distribution of molecular weight). Time invariant noises and baseline offsets were corrected before fitting. A maximum entropy regularization confidence level of 0.95 was used in all the size distribution analyses.

Polarization Fluorescence Assay. Polarization fluorescence was measured at room temperature with a Fluorolog-3 spectrofluorometer (HORIBA). ENH-c2b was serially diluted in buffer containing 100 mM NaCl and 20 mM TrisHCl at pH 8.0. Fluorescence anisotropy was measured for each sample, and the G-factor was determined individually. Data were analyzed according to a simple monomer-dimer equilibrium model and fit with KaleidaGraph software. Polarization values (mA) for the completely monomeric and dimeric states were fit to be 12 and 251, respectively.

X-Ray Crystallography. ENH-c2b crystals were grown at room temperature in 1% w/v tryptone, 20% w/v polyethylene glycerol 3350, and 0.05 M HEPES sodium at pH 7.0 using hanging-drop diffusion. Needle-like crystals appeared within 1 week. The crystals were soaked in glycerol cryoprotectant and flash frozen by cold nitrogen stream. Diffraction data were collected at beamline BL13C1 at the National Synchrotron Radiation Research Center in Taiwan. The best diffraction data had a resolution of ~ 2.2 Å. However, due to X-ray overexposure, the overall data quality was not ideal, so the data were truncated to 3.5 Å for better refinement. Phases were obtained through molecular replacement using ENH (PDB code: 1ENH) as the searching model. Further refinement was done with PHENIX. The data statistics are listed in Table 3-3. Final coordinates were deposited in the Protein Data Bank with the PDB code 4NDL.

Solution NMR Experiments. All spectra were acquired at 310 K on a Bruker Avance III 800 spectrometer equipped with a 5 mm z-gradient TCI (^1H , ^{13}C , and ^{15}N) cryoprobe (Bruker, Karlsruhe, Germany). ENH-c2b with an extra C-terminal glycine and Strep•tag II (1.9 mM protein in 300 μL) was dissolved in 100 mM NaCl, 5 mM CaCl_2 , 10 mM DTT, 0.02% NaN_3 , 5% D_2O , and 20 mM NH_4OAc at pH 4.5 in a Shigemi NMR tube (Allison Park, Pa., USA). Assignment of main-chain and side-chain chemical shifts was based on ^1H - ^{15}N HSQC, ^1H - ^{13}C HSQC, CBCA(CO)NH, HNCACB, HNCO, HNCACO, HCCH-COSY, HCCH-TOCSY, HBHANH, HNHA(CO)NH, (H)CC(CO)NH, H(C)CCONH, HNHA, CACO, CON, and ^{15}N -TOCSY-HSQC experiments. NOE distance restraints were obtained from ^{15}N -edited NOESY, ^{13}C -edited NOESY (aliphatic), and ^{13}C -edited NOESY (aromatic) for intra-chain or inter-chain contacts. An asymmetrically labeled dimer was prepared by mixing 1:1 uniformly $^{13}\text{C}/^{15}\text{N}$ labeled and unlabeled ENH-c2b. This sample was used for the $^{12}\text{C}/^{14}\text{N}$ filtered ^{13}C -edited NOESY experiment (mixing time = 300 ms) in order to extract the inter-chain NOE restraints. ^1H chemical shifts were referenced to 4,4-dimethyl-4-silapentane-1-sulfonate (DSS) as the external standard. ^{15}N and ^{13}C chemical shifts were referenced using the consensus ratios of the zero-point frequencies at 310 K. Data were processed with

Topspin (Bruker) for Fourier transformations and analyzed with CCPN for chemical shift assignments.

Solution Structure Determination. TOLOS+ (ref) was used for ϕ/ψ restraints predicted by backbone chemical shifts. Backbone hydrogen-bond restraints were created between consecutive $i/i+4$ helical residues. ϕ/ψ restraints, hydrogen-bond restraints, and a set of partially manual NOE assignments were used as the initial input for ARIA2.3. Automated NOE cross-peak assignments and structure calculations were then applied iteratively by ARIA2.3. For regular NOESY experiments, every NOE cross-peak was treated ambiguously as an inter- or intramolecular restraint. For the $^{13}\text{C}/^{15}\text{N}$ filtered ^{13}C -edited NOESY, the NOE cross-peaks were treated as intermolecular restraints only. A C2-symmetry restraint energy term was included. A soft square potential was used in the simulated annealing protocol with automated determination of weights for NOE-derived restraints. The 7 highest scoring structures out of a total of 32 generated structures were chosen for every cycle to obtain assignment statistics. A total of 8 cycles were run and the 10 lowest-energy models were refined in explicit water at the end to obtain the final NMR ensemble. The data statistics are listed in Table 3-4.

References

1. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. *Proteins* 78:3111-3114.
2. Xu D, Tsai CJ, Nussinov R (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* 10:999-1012.
3. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1-9.
4. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci* 9:1-15.

5. Whitehead TA, Baker D, Fleishman SJ (2013) Computational design of novel protein binders and experimental affinity maturation. *Method Enzymol* 523:1-19.
6. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47:409-443.
7. Shifman JM, Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci USA* 100:13274-13279.
8. Havranek JJ, Harbury PB (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* 10:45-52.
9. Sia SK, Kim PS (2003) Protein grafting of an HIV-1-inhibiting epitope. *Proc Natl Acad Sci USA* 100:9756-9761.
10. Lewis SM, Kuhlman BA (2011) Anchored design of protein-protein interfaces. *Plos One* 6:e20872.
11. Fleishman SJ, *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332:816-821.
12. King NP, *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336:1171-1174.
13. Der BS, *et al.* (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J Am Chem Soc* 134:375-385.
14. Sammond DW, *et al.* (2011) Computational design of the sequence and structure of a protein-binding peptide. *J Am Chem Soc* 133:4190-4192.

15. Whitehead TA, *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* 30:543-548.
16. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using β -strand assembly. *Proc Natl Acad Sci USA* 108:20562-20567.
17. Stranges PB, Kuhlman B (2013) A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci* 22:74-82.
18. Fleishman SJ, *et al.* (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *J Mol Biol* 414:289-302.
19. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins* 53:708-719.
20. Harbury PB, Zhang T, Kim PS, Alber T (1993) A switch between 2-stranded, 3-stranded and 4-stranded coiled coils in Gcn4 leucine-zipper mutants. *Science* 262:1401-1407.
21. Grigoryan G, Reinke AW, Keating AE (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* 458:859-864.
22. Guharoy M, Chakrabarti P (2007) Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics* 23:1909-1918.
23. Woolfson DN (2005) The design of coiled-coil structures and assemblies. *Adv Protein Chem* 70:79-112.

24. Apgar JR, Gutwin KN, Keating AE (2008) Predicting helix orientation for coiled-coil dimers. *Proteins* 72:1048-1065.
25. Karanicolas J, *et al.* (2011) A de novo protein binding pair by computational design and directed evolution. *Mol Cell* 42:250-260.
26. Fraenkel E, Rould MA, Chambers KA, Pabo CO (1998) Engrailed homeodomain-DNA complex at 2.2 angstrom resolution: A detailed view of the interface and comparison with other engrailed structures. *J Mol Biol* 284:351-361.
27. Marshall SA, Mayo SL (2001) Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305:619-631.
28. Shah PS, *et al.* (2007) Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol* 372:1-6.
29. Beck DAC, Daggett V (2004) Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods* 34:112-120.
30. Privett HK, *et al.* (2012) Iterative approach to computational enzyme design. *Proc Natl Acad Sci USA* 109:3790-3795.
31. Marshall SA, Morgan CS, Mayo SL (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol* 316:189-199.
32. Huang P-S, Love JJ, Mayo SL (2005) Adaptation of a fast Fourier transform-based docking algorithm for protein design. *J Comput Chem* 26:1222-1232.
33. Khare SD, Fleishman SJ (2013) Emerging themes in the computational design of novel enzymes and protein-protein interfaces. *FEBS Lett* 587:1147-1154.

34. Röthlisberger D, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453:190-195.
35. Jiang L, *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319:1387-1391.
36. Kiss G, Röthlisberger D, Baker D, Houk KN (2010) Evaluation and ranking of enzyme designs. *Protein Sci* 19:1760-1773.
37. Oshea EK, Klemm JD, Kim PS, Alber T (1991) X-Ray structure of the Gcn4 leucine zipper, a 2-stranded, parallel coiled coil. *Science* 254:539-544.
38. Dahiyat BI, Mayo SL (1996) Protein design automation. *Protein Sci* 5:895-903.
39. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372:774-797.
40. Procko E, *et al.* (2013) Computational design of a protein-based enzyme inhibitor. *J Mol Biol* 425:3563-3575.
41. Murphy PM, Bolduc JM, Gallaher JL, Stoddard BL, Baker D (2009) Alteration of enzyme specificity by computational loop remodeling and design. *Proc Natl Acad Sci USA* 106:9215-9220.
42. Mandell DJ, Coutsiar EA, Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* 6:551-552.
43. Hu XZ, Wang HC, Ke HM, Kuhlman B (2007) High-resolution design of a protein loop. *Proc Natl Acad Sci USA* 104:17668-17673.

44. Malakauskas SM, Mayo SL (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* 5:470-475.
45. Korkegian A, Black ME, Baker D, Stoddard BL (2005) Computational thermostabilization of an enzyme. *Science* 308:857-860.

Table 3-1. Library design of the homodimer interface

	9	10	13	14	16	17	18	25	28	32
Library	K	AT	L	A	FY	FV	DFVY	AW	FY	R

Computational library design for homodimerization. Interfacial residues shown above were chosen for sequence optimization. The library size was set to 128 members.

Table 3-2. Sequences of wild-type ENH, ENH-c2a, NC3-Ncap, and ENH-c2b



		id*	Tm†
ENH (WT)	TAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI	-	49
ENH-c2a	-----KA-DL--YF-----W--Y---R-----	82%	n/a
NC3-Ncap	-E--E--KR--DE--RRD-R---E--RD--QK-----E--ER--RR-EQQ-	55%	88
ENH-c2b	-E--E--KKA-DLA-YFD-R--PEW-RY--QR-----E--ER--RR-EQQ-	47%	62

The three “coils” at the top show the location of the three helices in the ENH wild-type (WT) fold.

*id is the sequence identity compared to ENH.

†Tm is the melting temperature (°C).

Table 3-3. Data collection and refinement statistics for the crystal structure ENH-c2b, with an extra 21-residue tag at N-terminus (PDB code: 4NDL)

Statistics	Value
Data collection	
Space group	C222 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	87.6, 167.8, 29.7
α , β , γ (°)	90.0, 90.0, 90.0
Resolution (Å)	25.9–2.2
R _{merge}	0.147
I/ σ	8.0
Completeness, %	94.1
Multiplicity	5.3
Refinement	
Resolution (Å)	24.6–3.5
Number of reflections	5697
R _{work} /R _{free} (%)	31/32
Number of molecules in asymmetric unit	3
Number of atoms	2024
Protein	2023
Water	1
B factors (Å ²)	27.3
Protein	27.4
Water	17.1
R.m.s. deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.77
Ramachandron map analysis	
Most favored regions (%)	96.9%
Additional allowed regions (%)	3.1%
Disallowed regions	0.0%

Table 3-4. NMR statistics for the structure ENH-c2b-Strep (PDB code: 2MG4)

NMR structure statistics	
Summary of restraints	
Total NOE distance restraints	1347
Intra-molecular unambiguous	1134
Intra-molecular ambiguous	186
Inter-molecular unambiguous	26
Inter-molecular ambiguous	1
Hydrogen bonds	42
Dihedral angle restraints (ϕ/ψ)	82/82
r.m.s. deviation from restraints	
NOE restraints (Å)	0.061 ± 0.021
H-bond restraints (Å)	0.026 ± 0.010
Dihedral restraints ($^\circ$)	1.16 ± 0.26
r.m.s. deviation from idealized geometry	
Bonds (Å)	0.0051 ± 0.0002
Angles ($^\circ$)	0.72 ± 0.03
Improper ($^\circ$)	1.75 ± 0.17
Coordinate precision r.m.s.d. (Å)	
backbone, secondary structure	0.64 ± 0.23
heavy atoms, secondary structure	1.19 ± 0.19
backbone, all	1.40 ± 0.40
heavy atoms, all	2.20 ± 0.34
Ensemble Ramachandran statistics (%)	
residues in most-favored region	87.9
additionally allowed region	7.6
generally allowed region	4.5
disallowed region	0.0

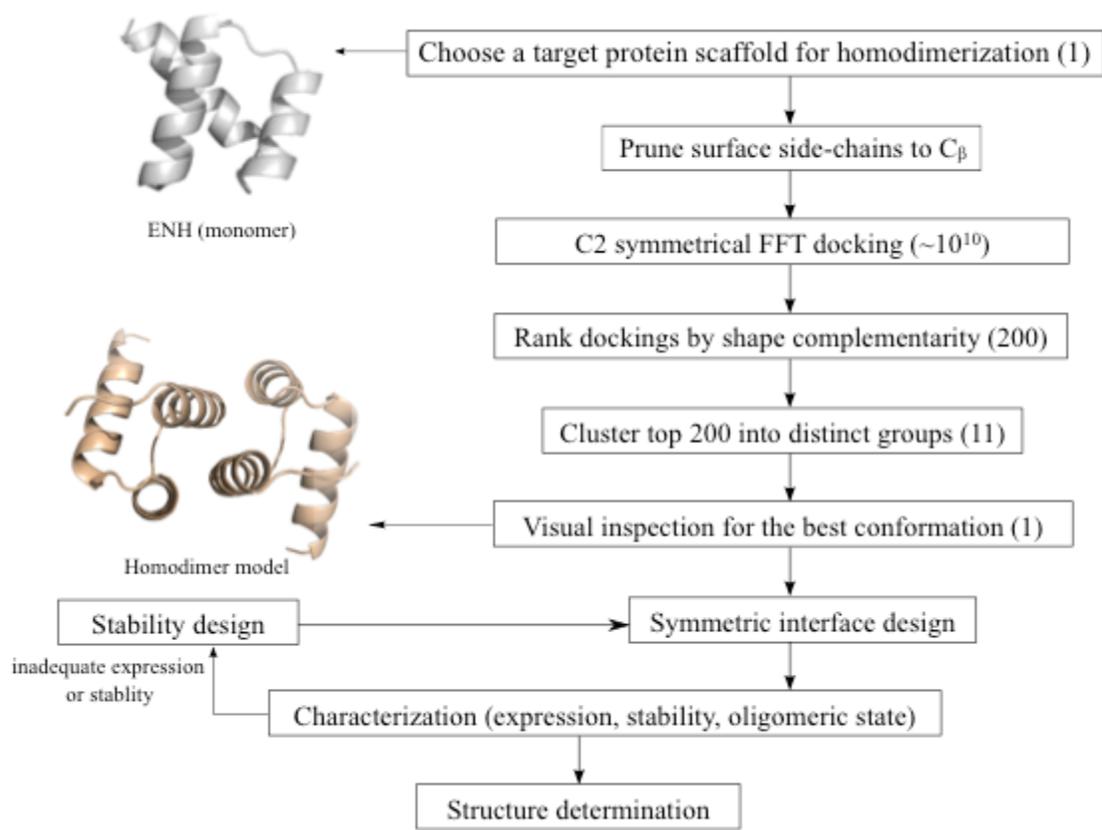


Fig. 3-1. Steps used to design a C2-symmetrical homodimer. The initial scaffold (ENH) used for docking is shown in silver, and the homodimer model used for all interface designs is shown in bronze. The number of models created in each step is given in parenthesis.

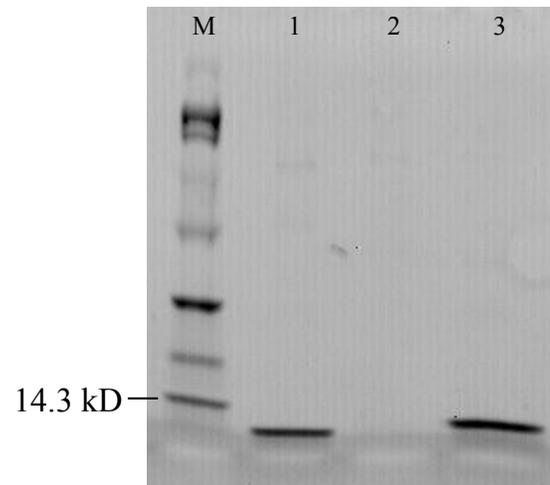


Fig. 3-2. SDS-PAGE of purified proteins from soluble fractions. The labels are M: marker, 1: ENH, 2: ENH-c2a, and 3: ENH-c2b. The absence of band in lane 2 indicates that ENH-c2a is not expressed in soluble fraction.

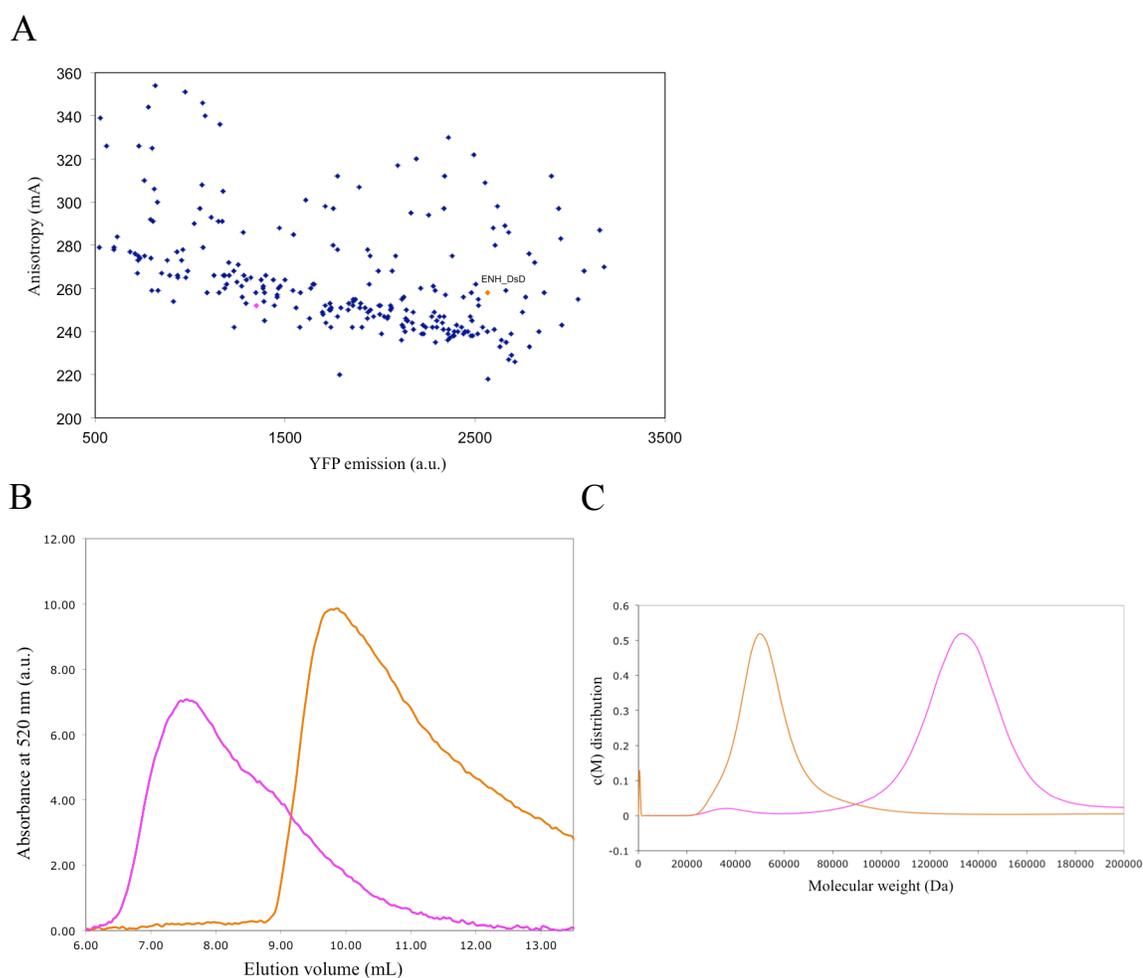


Fig. 3-3. Experimental characterizations of the computational library design. (A) Homo-FRET assay of the library members. Each dot represents one member in the library. Because of the homo-FRET effect, members that have lower anisotropic values are likely to have stronger dimer affinities or higher oligomeric states. Two representative members, shown in orange (named as ENH_DsD) and magenta were chosen for characterizations in B and C. (B) Size-exclusion chromatography showed that ENH_DsD (orange curve) is likely a dimer (compared to protein standards) and magenta variant formed a higher-order oligomer. (C) Sedimentation velocity experiments showed that ENH_DsD is under a fast equilibrium between monomer and dimer states (orange curve); magenta variant is a tetramer (magenta curve). The monomer MW is 34.4 kD.

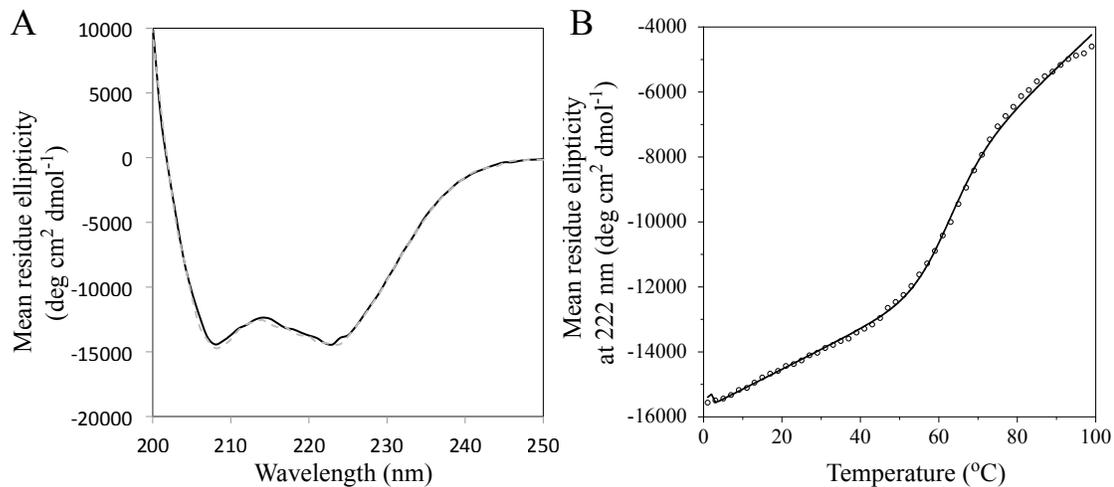


Fig. 3-4. CD spectroscopy shows that ENH-c2b is a fully refoldable helical protein. (A) CD spectrum of ENH-c2b at room temperature. Solid line: before thermal denaturation; dashed line: after thermal denaturation. (B) Thermal denaturation curve measured at 222 nm. Circles: experimental data; line: fitted curve obtained using a two-state transition model.

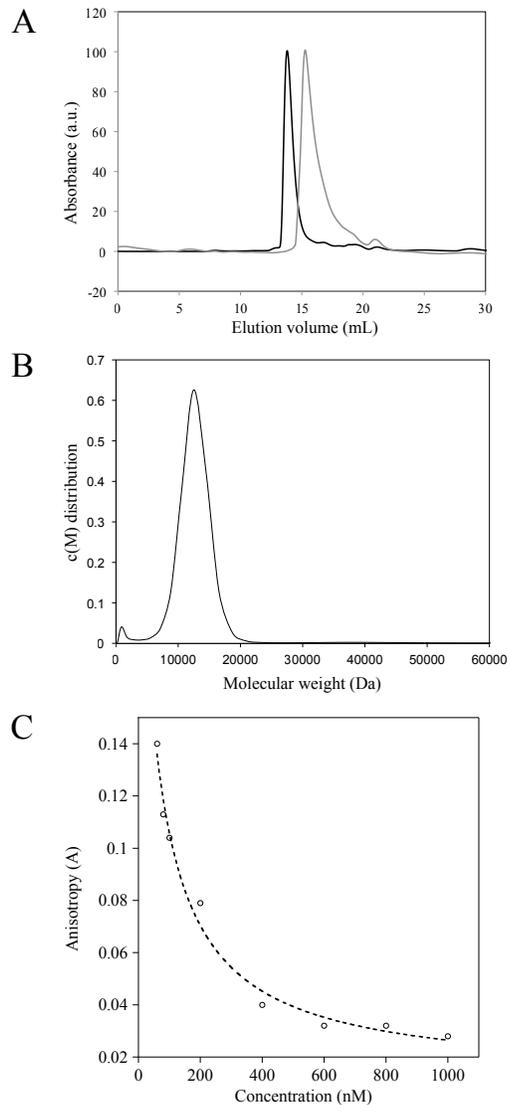


Fig. 3-5. Characterization of oligomeric state reveals that ENH-c2b is a homodimer with $K_d \sim 130$ nM. (A) Size-exclusion chromatography at 280 nm; experiments were done using loading protein concentrations of ~ 500 μM (black) and 10 μM (gray), indicating dimeric and monomer states, respectively. Curves were normalized to 100 to facilitate comparison. (B) Sedimentation velocity experiment at 5 μM ; curve shows data fit using the c(M) model. The major peak centered at 12429 Da indicates a dimer (MW of monomer = 6531). (C) Tryptophan homo-FRET assay shows $K_d \sim 130$ nM. Circles: experimental data; dashed line: fitted curve obtained using a monomer-dimer equilibrium model.

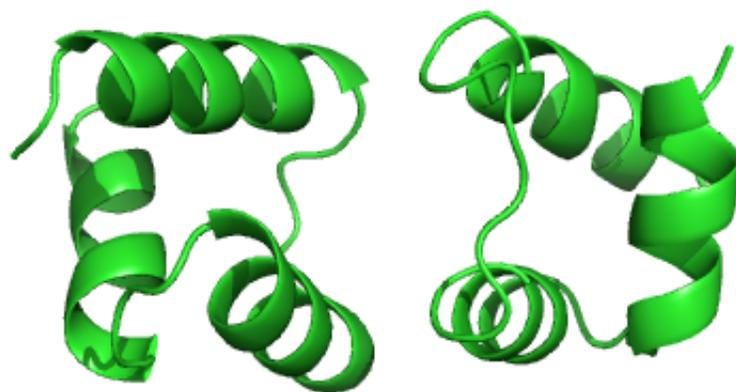


Fig. 3-6. Crystal structure of ENH-c2b with the long N-terminal tag “MGSSHHHHHSSGLVPRGSHM” (PDB code: 4NDL). The two chains in proximity make a crystal contact with helix-1 and helix-2. Inspection evaluated by PISA reveals that this contact belongs to a crystal packing rather than a biological interface, mainly because of its small area (699 Å², the average interface area for a homodimer is 2740 ± 1240 Å²).

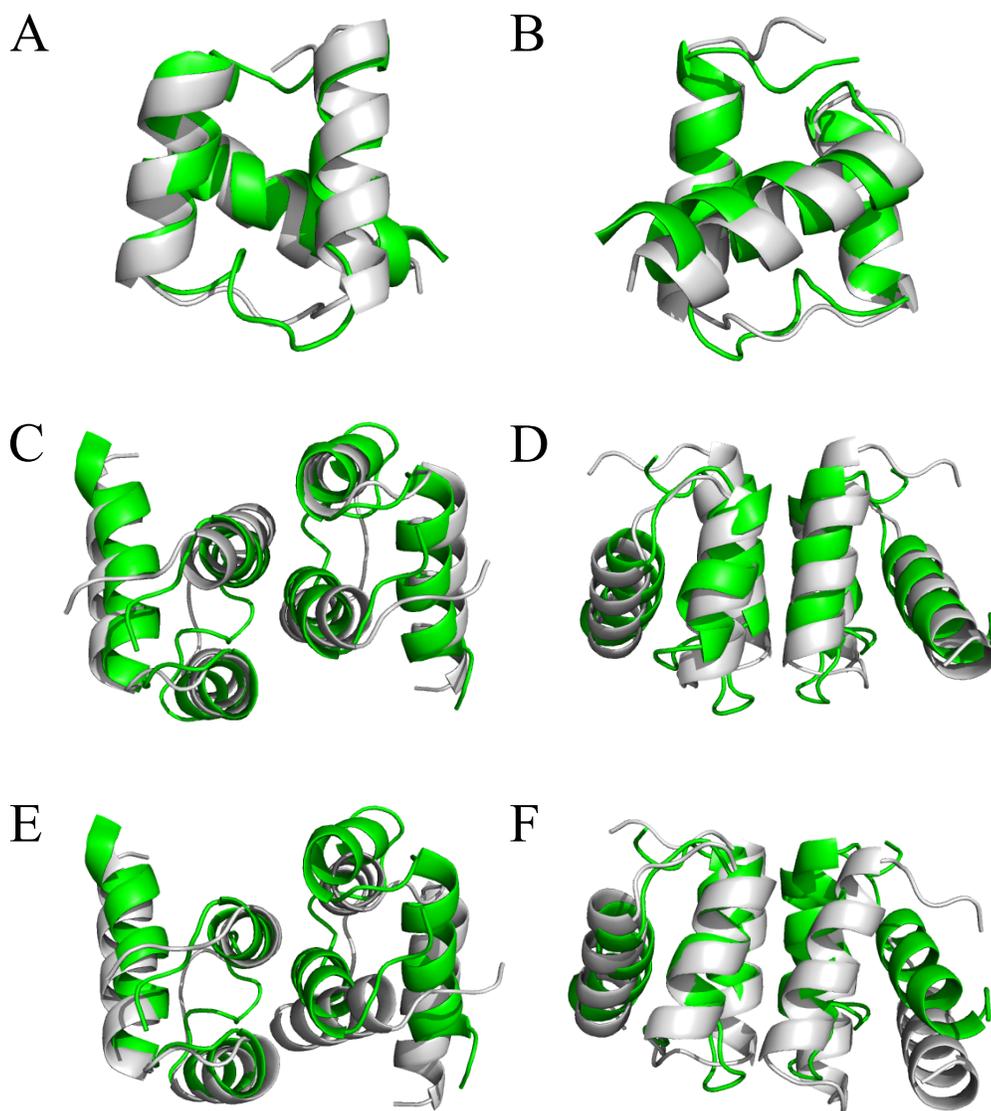


Fig. 3-7. Solution NMR structure of ENH-c2b (green) superimposed with design model structure (gray), viewed from different orientations. (A-B): superposition of just one chain; (C-D): superposition of the entire dimer structure; (E-F): superposition of left chain showing entire dimer structure.

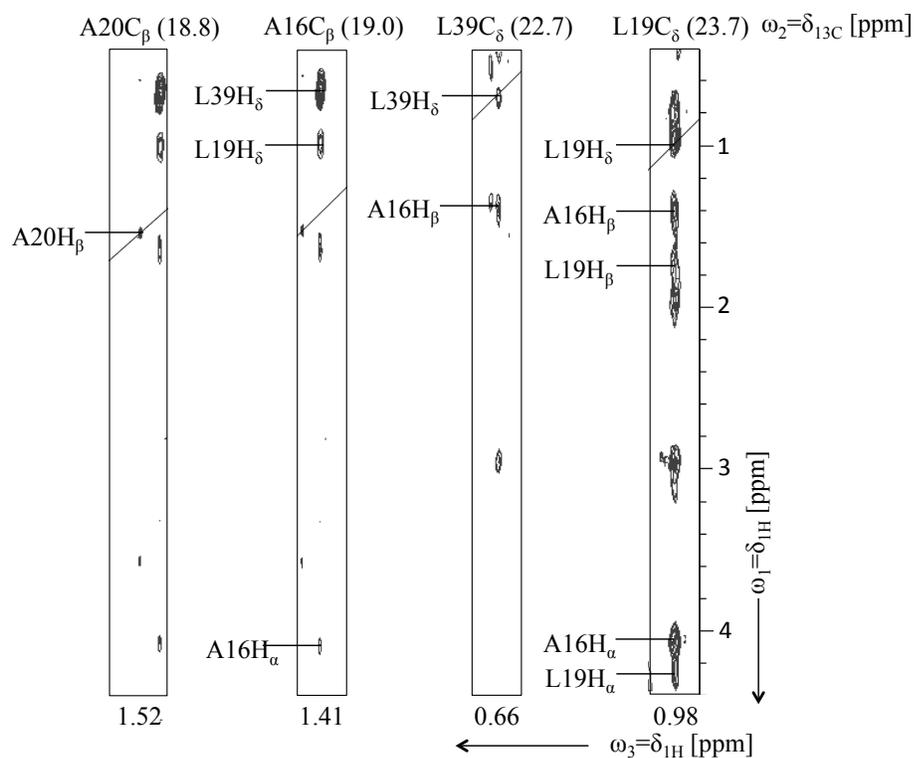


Fig. 3-8. NMR spectrum showing intermolecular NOE restraints obtained by $^{12}\text{C}/^{14}\text{N}$ filtered ^{13}C -edited NOESY experiment. Contour plots of $[\omega_1(^1\text{H}), \omega_3(^1\text{H})]$ -strips of Ala16 C_β , Ala20 C_β , Leu19 δ and Leu39 C_δ are shown. Chemical shifts indicated on top and bottom correspond to $\omega_2(^{13}\text{C})$ and $\omega_3(^1\text{H})$ dimensions, respectively. For clarity, only the aliphatic region in the $\omega_1(^1\text{H})$ dimension is shown. Unambiguous restraints identified for Ala16, Ala20, Leu19, and Leu39 residues are labeled.

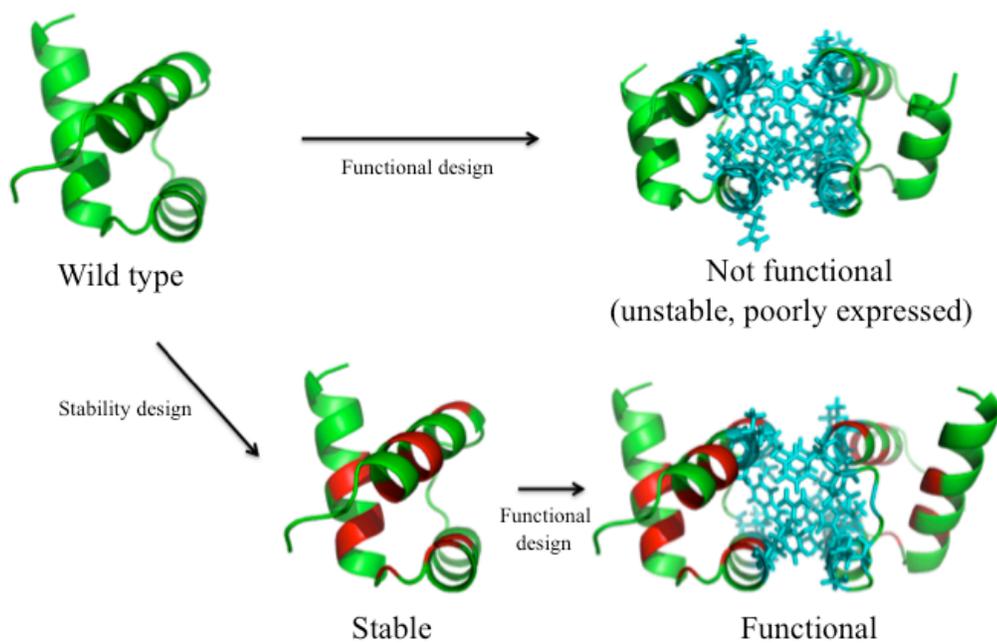


Fig. 3-9. Two-step design strategy for the functional design of homodimers. A single-step design for functionality often leads to an unstable protein. Stability design prior to functional design can solve this problem. In this work, single-step functional design of ENH led to an unstable protein that did not express solubly, referred to as ENH-c2a. However, doing a stability design (NC3-NCap) prior to functional design led to a stable homodimer, ENH-c2b. Structures shown are computational design models.

Chapter 4

Computational Design of Self-assembling

Protein-DNA Nanowires

Abstract

Computation protein design (CPD) has been successfully used to create various kinds of functional proteins, including enzymes, protein oligomers, and ligand binders. The ability to rationally design molecular self-assembly using biological macromolecules is of particular interest because of its versatility for applications in biotechnology and medicine. Sophisticated single-component nanostructures composed of nucleic acids or proteins have been demonstrated, but despite these successes, the development of hybrid self-assemblies of nucleic acids and proteins via non-covalent interactions remains elusive. Here, we used CPD to create a protein-DNA complex that can self-assemble into nanowires. To achieve this, a homodimerization domain was engineered into the transcription factor engrailed homeodomain (ENH) so that it could bind to two DNA molecules on the two opposite sides. The homodimerization domain was designed *de novo*, whereas ENH's native DNA-binding domain was exploited to bind a specific double-stranded DNA (dsDNA) motif. When dsDNA fragments containing multiple copies of this motif were combined with the engineered ENH homodimer, the two components polymerized via non-covalent interactions to form nanoparticles. Particle formation was completely inhibited by adding single binding-site dsDNA, confirming that self-assembly occurred via the specified protein-DNA interactions. Furthermore, if the dsDNA fragment contains only two binding motifs on the exact opposite sides, the protein-DNA self-assembly led to nanowires as visualized by atomic force microscopy. The diameter of the nanowire is about 10 nm, which is consistent with the length of the dsDNA fragment. The length of the nanowire is up to 300 nm. The protein-DNA co-crystal structure confirmed that the nanowire is formed via the designed interactions. To the best of our knowledge, this is the first example of DNA-protein nanomaterial that is self-assembled via pure non-covalent interactions.

Introduction

Biomolecular self-assembly is an essential process for living organisms as well as a bottom-up approach for bio-nanotechnology (1). The ability of scientists to control the self-assembly of macromolecules, including nucleic acids and proteins, has progressed significantly in the past decade. DNA has been engineered into “origami,” which has led to the creation of seemingly arbitrary two- and three-dimensional nanomaterials (2). DNA self-assemblies have been shown to have many useful applications ranging from drug delivery (3) to molecular computing (4). The design of protein self-assemblies, in contrast, has proved to be more challenging, owing in part to our incomplete understanding of protein folding and protein-protein interactions. Nevertheless, a single polypeptide chain comprised of a concatenation of coil-coiled motifs was recently designed to fold into a defined tetrahedron nanostructure (5). Moreover, highly symmetrical tetrahedral and octahedral oligomeric protein cages have been computationally designed with atomic level accuracy (6).

The functionality of these nucleic acid or protein materials, however, has been limited by the physical/chemical nature of a single building block. To develop integrated materials with multiple cooperative functions, it is advantageous to use two or more kinds of building blocks for self-assembly. Hybrid materials in which at least one of the components has a characteristic length in the nanometer range often exhibit superior properties over single-component materials. Self-assemblies of protein and DNA, for example, have been engineered for the highly sensitive detection of proteins (7) and for the programmed positioning of multienzyme cascades (8). To the best of our knowledge, all protein-DNA hybrid materials engineered to date use chemical conjugations to link the DNA and protein molecules either covalently or through a streptavidin-biotin interaction. Depending on chemical conjugation for self-assembly, however, complicates building-block synthesis considerably and can lead to heterogeneous labeling problems.

Designing a protein-DNA self-assembly driven by non-covalent interactions holds great promise, but is hindered by our limited understanding of protein-protein and protein-DNA

interactions. Recent work has demonstrated the engineering of novel protein-protein interfaces (9, 10) and the alteration of protein-DNA binding specificity (11). The creation of a protein-DNA self-assembly, however, raises the additional complex problem of successfully arranging these binding interfaces to allow complex formation. Any self-assembly with a homogeneous composition requires that at least two binding sites or domains exist on every building block. Natural proteins frequently use a multi-domain approach for complex tasks. For example, many transcription factors use dual DNA-binding and oligomeric activating domains to achieve precise control of gene regulation (12).

Computational protein design (CPD) is a powerful tool that has recently been applied to create proteins with new functionality. In the last few years, enzymes have been designed *de novo* to catalyze non-natural enzymatic reactions (13-16), and protein interactions have been designed to generate novel heterodimers (10), homodimers (9), and ligand binders (17) with K_d values as low as the nanomolar range. Most functional protein designs to date have focused on building a desired functionality into a protein, while the protein's native function is ignored or even lost. In a successful recent counter-example, King et al. designed protein nanomaterials with atomic level accuracy using native trimers as building blocks, computationally adding a new protein-protein interface to a protein that had an existing trimerization interface (6). These two interfaces together allow the protein to self-assemble into a highly symmetrical cage, emphasizing the power of having two cooperative functions in a single protein.

Here, we propose an analogous strategy to make self-assembled DNA-protein nano-objects via non-covalent interactions. The strategy relies on the judicious design of a dual-function protein with two independent binding domains: a homodimerization domain and a DNA-binding domain. This designed protein should form a homodimer, with each constituent monomer binding a specific fragment of double-stranded DNA (dsDNA). If the dsDNA has multiple specific binding sites for the protein, we expect to see spontaneous self-assembly of a protein-DNA complex. We first test the homodimerization and DNA-binding functionalities of our designed protein and subsequently

show that the designed protein self-assembles with its target dsDNA to form nanoparticles. Moreover, if we designed the dsDNA with exact two binding sites located on the opposite sides, the protein-DNA self-assembly formed nanowires. Finally, we demonstrate the functionality of the protein-DNA nanoparticles and discuss potential applications and further studies.

Results

Scaffold Selection. In selecting our scaffold protein, we required that it have two clear binding domains, one for DNA binding and one for protein-protein homodimerization. Additionally, the two domains had to be adequately removed from each other structurally so as not to conflict with binding and to permit large-scale complex assembly. Because the *de novo* design of DNA-binding functionality into a protein has not yet been achieved, we decided to use an existing DNA-binding interface in this study. The helix-turn-helix motif is one of the most abundant DNA-binding motifs in natural proteins (18). Its principal mechanism of action is the interaction of a positively-charged helix with the major groove of dsDNA in a sequence-specific manner. We chose engrailed homeodomain (ENH) as our scaffold for the following reasons: (1) a previous study showed that a single-helix peptide isolated from ENH can bind a target dsDNA motif (TAATNN) as tightly as the full-length protein (K_d in nM range) (19); (2) ENH has been intensely studied using computational tools, so highly stable variants as well as full-sequence redesigns have been generated (20, 21); and (3) ENH is a three-helix protein, providing a surface for homodimer interface design that is structurally separated from the DNA-binding functionality. The surface targeted for homodimerization comprises helices 1 and 2, which are anti-parallel to each other and opposite DNA-binding helix 3. This large, flat surface shares no residues or secondary structure with the DNA-binding domain. In addition, this potential homodimerization domain is on the opposite side of the DNA-binding domain. The spatial arrangement of these two binding sites can possibly be used to self-assemble a linear structure, like nanowires. Thus, we expected ENH would serve as an excellent scaffold for the design of a protein-DNA self-assembling nanostructure.

Design Protocol. Figure 4-1 illustrates our protein-DNA nanomaterial design strategy. Using the *Drosophila melanogaster* ENH crystal structure (PDB code: 1ENH) (22) as our docking subunit (Fig. 4-1A), we performed fast Fourier transform-based docking to generate C2 symmetrical homodimer models (23). The best model exhibited parallel intermolecular helical packing between helices 1 and 2 of each of the ENH monomers. CPD was then used to design the interface residues of the docked model to minimize the free energy of the intermolecular side-chain interactions (Fig. 4-1B). Early design variants were characterized and iteratively improved with the use of a molecular dynamics screening protocol (24, 25). We named the final designed variant dualENH because it has dual functionality: it can both homodimerize and bind dsDNA. A dualENH homodimer serves as the protein building block for nanomaterial assembly because it has two binding sites for dsDNA (one on each of two opposite faces of the homodimer), as shown in an aligned model (Fig. 4-1C). The second designed component of the nanomaterials is a dsDNA building block with protein binding sites variously placed along the double helix (Fig. 4-1D and Fig. 4-2A). By tuning the positioning of binding sites on the dsDNA and then simply mixing the two designed components (DNA and protein) together, we were able to achieve co-assembly of both irregularly shaped particles of protein and DNA (Fig. 4-2B) and well-ordered protein-DNA nanowires (Fig. 4-1E).

Biophysical characterization of dualENH. A synthetic gene encoding the dualENH sequence was constructed with a C-terminal (His)₆-tag, cloned into an expression plasmid, and transformed into *Escherichia coli* cells. Expression of dualENH at 37 °C produced >10 mg of soluble protein per liter of *E. coli* culture. Far-UV circular dichroism (CD) spectroscopy showed that dualENH is a fully refoldable, entirely α -helical protein (Fig. 4-3A). The CD curve for dualENH is very similar to that of wild-type ENH, suggesting a high degree of structural similarity. Thermal denaturation monitored by CD showed that dualENH has a melting temperature (T_m) of 59 °C (Fig. 4-3B), which makes it more stable than wild-type ENH (T_m = 49 °C).

Size-exclusion chromatography and analytical ultracentrifugation were used to determine the oligomeric state of dualENH. Different initial concentrations of dualENH were run over a

Superdex 75 size-exclusion column. As the protein concentration was reduced, the peak elution volume gradually moved from an earlier position (~14 ml) to a later position (~19 ml) (Fig. 4-4A), indicating that dualENH is present in different oligomeric states at different concentrations. To determine its oligomeric state explicitly, a sedimentation velocity experiment was run, which showed that the principal dualENH species present at 40 mM is a homodimer, with a very small amount of higher-order oligomers (Fig. 4-4B).

We used fluorescence polarization to determine whether dualENH binds dsDNA probes strongly and specifically. The polarization will increase if a dsDNA probe is bound by dualENH due to a reduction in the tumbling rate of the larger complex. We used a fluorescein-labeled dsDNA probe containing the ENH binding motif TAATTA (probe-1) that had previously been used in wild-type ENH binding studies (26). The polarization of a 25 nM solution of probe-1 increased from its intrinsic value of ~140 mP to a plateau of 210 mP as the concentration of dualENH was increased from 0 nM to 100 nM (Fig. 4-4D). The same experiment run with wild-type ENH showed a very similar trend and polarization values (Fig. 4-5A), indicating that dualENH and wild-type ENH have similar binding affinities to probe-1. To test the binding specificity, we designed probe-2, which is identical to probe-1 except for a single mutation to its binding motif (TA[C]TTA). Compared to probe-1, a weaker response was observed between probe-2 and dualENH: polarization increased to only 176 mP at 100 nM dualENH and did not plateau until the concentration of dualENH reached 10 μ M (Fig. 4-4D). An additional probe that had previously been used as a negative control (27) for wild-type ENH binding was also tested. As expected, neither dualENH nor wild-type ENH showed observable binding to this probe at 100 nM of protein (Fig. 4-5B). Together, these data show that dualENH binds strongly and specifically to the wild-type ENH binding motif, TAATTA.

We next sought to confirm that each dualENH homodimer could bind two dsDNA fragments as illustrated in Fig. 4-1C. Using a Förster resonance energy transfer (FRET)-based experiment, a 15-nt dsDNA sequence (TAA)₅ was labeled with Cy3 or Cy5 dye to serve as a FRET donor or acceptor, respectively. We mixed the two labeled (TAA)₅ probes with dualENH and observed a strong FRET

signal as shown in Fig. 4-4E indicating that the two pieces of dsDNA were brought within Förster distance by dualENH. These experiments indicate that the DNA-binding domain and the homodimerization domain of dualENH are structurally independent, and that they can function cooperatively.

Co-assembly of Protein-DNA materials. We then devised an experiment to determine whether these two functions could act synergistically to self-assemble protein-DNA nano-objects with specific dsDNA fragments. We sought to observe the protein-DNA self-assembly using fluorescence microscopy. Figure 4-6A shows that nanoparticles were formed immediately after 5 μM dualENH was mixed with 2 μM (TAA)₅. The particles were irregularly shaped with diameters of up to several microns. The irregularity of shape was expected because (TAA)₅ has four ENH binding sites (TAATAA) that each face in a different direction off of the dsDNA helix, which causes particle growth to occur in a random branching pattern (Fig. 4-2B). The particles are invisible under bright-field microscopy (Fig. 4-7B) because of the transparency of protein and DNA to visible light. A control experiment showed that a solution of (TAA)₅ by itself did not form any particles (Fig. 4-7C). When lower concentrations of dualENH (500 nM) and (TAA)₅ (200 nM) were used, a smaller and more uniform particle distribution was observed (Fig. 4-7D). Further decreasing protein concentration (< 200 nM) significantly reduced the number of particles formed (Data Fig. 4-7E). This reduction may be due to dissociation of the homodimer at low concentrations. To confirm that nanoparticle formation occurs via the proposed mechanism (Fig. 4-2B), we designed a particle inhibition experiment using dsDNA with a single binding site as the inhibitor. The single binding site on these dsDNA should terminate particle growth. Preincubation of 500 nM dualENH with only trace amounts of single-binding-site dsDNA (5 nM) abolished particle formation completely when 200 nM (TAA)₅ was added (Fig. 4-7F).

To form a linear protein-DNA co-assembly as illustrated in Fig. 4-1E, the dsDNA building block must have two protein binding sites about 180° apart on the dsDNA double helix (Fig. 4-1D). We designed a 25-nt dsDNA molecule with an 11-nt binding motif (TAATTTAATTT, named

motif-11) that contains two ENH binding motifs (TAATTT) facing in opposite directions off of the helix. At the same protein and DNA concentrations used in the earlier particle-forming experiment (5 μ M and 2 μ M, respectively) (Fig. 4-6A), dsDNA containing motif-11 and dualENH formed much smaller and more uniform particles, with none growing greater in size than the diffraction limit (submicron) (Fig. 4-6B). The reduced particle size may be a result of fewer protein binding sites on the DNA building block (2 vs. 4). Fewer protein binding sites would decrease the entropy of self-assembly and increase the chance of binding-site poisoning. We used atomic force microscopy (AFM) to study the topology of the protein-DNA co-assembly formed with dualENH and the two-binding-site dsDNA. Nanowire structures were clearly observed with a width of \sim 15 nm and a length of up to \sim 300 nm (Fig. 4-8A, B). In accordance with the design model, the observed width of the nanowires (\sim 15 nm) is consistent with the length of the dsDNA (\sim 9 nm), considering that AFM usually overestimates the length in the x-y plane due to the size of the tip. The height of the nanowire is \sim 1.0 nm (Fig. 4-8C), which is on the order of the diameter of a dsDNA fragment (\sim 2 nm); the decreased height could be due to compression by the hard AFM tip ($k = 3$ N/m).

We solved the co-crystal structure of dualENH and a dsDNA probe containing motif-11 (Fig. 4-9). The co-crystal structure confirms the dual functionality of dualENH: (1) dualENH uses helix-3 to bind dsDNA just as wild-type ENH does, and (2) dualENH uses the surfaces of helix-1 and helix-2 to form a homodimer (Fig. 4-9A). However, the co-crystal structure reveals two homodimer configurations of dualENH that differ slightly from each other, as seen by their backbone root mean square deviation (RMSD) of 4.0 \AA (Fig. 4-9B). This unexpected result might be caused by crystal packing forces, especially as each dsDNA molecule in the crystal forms a superhelix (see end-to-end packing of dsDNA fragments in Fig. 4-10A). Given this observation, we cannot conclude that either of the observed dualENH dimers in the crystal structure reflects the predominant dimer structure in solution. The two dualENH dimer crystal structures have backbone RMSDs to the design model of 3.8 \AA and 3.9 \AA , respectively (see Fig. 4-10B, C). The co-crystal structure also confirms that dualENH binds the dsDNA with its designed 11-nt binding motif;

however, it also reveals two configurations of protein-DNA binding. One configuration is consistent with our design model in which two dualENH molecules bind to the 11-nt motif exactly opposite each other on the dsDNA double helix (Fig. 4-9C). The other configuration, however, features one of the dualENH bindings in an inverted orientation; i.e., it binds to the reverse complementary sequence (AAATTA) of the optimal binding motif (TAATTT) (Fig. 4-9D). This suboptimal binding has also been seen in other ENH crystal structures(28), presumably due to the high concentrations of protein and DNA used for crystallization. Because of this alternate protein-DNA binding configuration, the nanowire in the co-crystal structure is slightly kinked; nevertheless, an infinitely repeated protein-DNA nanowire is observed (Fig. 4-9E).

We used CPD to design a protein-DNA nanomaterial, whose co-assembly is purely driven by non-covalent interactions. Unlike assemblies that rely on chemical conjugations, non-covalent self-assemblies can be tuned by altering the reaction conditions (e.g., temperature, pH, salt concentration). Indeed, our protein-DNA nanostructures will not form in high concentrations of salt because the protein-DNA electrostatic interaction is shielded (see Fig. 4-11). Very recently, King et al. developed a co-assembling system for protein nanomaterials, in which two different proteins are required for assembly (27). Co-assembling systems provide several advantages over single-component self-assemblies, including better control over the localization and timing of assembly, and greater functional and structural versatility of the assembly, as each component can confer unique attributes, especially in the case of hybrid materials such as those designed here. The protein-DNA nanostructures we designed could be further functionalized with the incorporation of engineered DNA structures, such as DNA origami or DNA aptamers. Furthermore, dualENH could be fused to peptide tags for antibody recognition or used for the specific attachment of organic and inorganic materials(29). We anticipate that protein-DNA co-assembly will open up many new possibilities for advanced biomaterial designs.

Discussion

CPD has been used to create a wide variety of functional proteins, from enzymes and homo- and hetero-dimers to small molecule binders and protein nanomaterials. Here, we used CPD to design the first protein-DNA nanoparticle whose self-assembly is driven by non-covalent interactions. Our results demonstrate that self-assembly can be achieved by engineering a dual-function protein that can both homodimerize and bind to a specific DNA fragment. We exploited the naturally occurring protein-DNA interaction of ENH and created a novel homodimer interaction via CPD. The homodimer and DNA-binding interactions occur on two non-overlapping interfaces, which allows them to function independently and concurrently. We designed an ENH homodimer that provides two binding sites for dsDNA, chose a dsDNA molecule with two opposite protein binding sites, and showed that self-assembly of protein-DNA nanowires occurs spontaneously. To the best of our knowledge, this is the first example of protein-DNA self-assembling nanomaterial formed via non-covalent interactions.

This multiple binding site approach has been used previously to drive the self-assemblies of protein-protein (30), protein-DNA (31), and protein-inorganic (32) nano-objects, but many of these nanomaterials required engineering of covalent attachments. Frequently, naturally occurring interactions have been leveraged to arrive at these self-assemblies, such as the homo-tetramerization of streptavidin and the streptavidin-biotin interaction. In these cases, chemical conjugations are necessary for the biotinylation of building blocks, which illustrates a disadvantage of this method. Protein biotinylation is usually targeted to specific amino acid side-chains, such as lysine or cysteine. If a protein has multiple exposed lysines or cysteines, the site-specificity of the labeling cannot be controlled. This results in heterogeneous labeling, which allows the haphazard formation of multiple types of assemblies created via different streptavidin-biotin interactions. In contrast, our self-assembled nanoparticles do not require any chemical conjugations, as they are driven by non-covalent interactions. More importantly, the binding sites for both protein-protein homodimerization and protein-DNA binding are specific and well-defined.

Materials and Methods

Protein docking and computational design. The details of protein docking and computational design were described elsewhere (25). Briefly, the ENH crystal structure (PDB code: 1ENH)(22) was used as the scaffold for homodimerization. A symmetrical docking program based on a fast Fourier transform (FFT) algorithm was applied (23), and one high-scoring model was selected for computational designs. Homebuilt CPD software was used for symmetry-constrained homodimer designs. Sequence optimization was performed using an improved version of FASTER (33) using a rotamer library based on the backbone-dependent library of Dunbrack and Karplus (34). The sequences of dualENH and wild-type ENH are listed in Table 4-1.

Construct preparation, protein expression, and purification. Oligonucleotides (Integrated DNA Technologies) containing ~20 bp overlapping segments were assembled via a modified Stemmer polymerase chain reaction (PCR) method (35) using KOD Hot Start Polymerase (Novagen) to generate genes for wild-type ENH and dualENH. A His₆ tag and a Gly-Ser linker (GGSGG) were added to the C terminus. Proteins were expressed using BL21 DE3 cells transformed by pET plasmids with 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) in standard Luria Broth (LB) at 37 °C. Proteins were purified from supernatant of lysed cells using affinity chromatography (Ni²⁺-NTA, Qiagen) followed by size-exclusion chromatography (Superdex 75, Amersham Pharmacia). Expression of dualENH at 37 °C produced >10 mg of soluble protein per liter of *E. coli* culture.

Circular dichroism spectroscopy. CD studies were performed on an Aviv 62A DS spectropolarimeter equipped with a thermoelectric temperature controller. Samples were prepared in 100 mM sodium chloride and 20 mM sodium phosphate buffer at pH 7.5. Wavelength scans and temperature denaturations were carried out in cuvettes with a 0.1 cm pathlength at a protein concentration of ~10 μ M. Three wavelength scans were performed at 25 °C for each sample and averaged. The thermal denaturation curve was collected at 222 nm from 0 °C to 99 °C, sampling

every 1 °C separated by 2 min equilibration times (signal averaging time was 1 sec). The refolding curve was collected after the thermal denaturation experiment using the same sample.

Analytical ultracentrifugation. dualENH was analyzed on an XL-1 analytical ultracentrifuge equipped with an AnTi60 rotor (Beckman Coulter). Two-channel epon-filled centerpieces were used for the sedimentation velocity experiment. Cells were torqued to 130 lb-inch and run at 60,000 rpm. Data were acquired at 280 nm at 20°C in continuous mode. Data were first fit to the c(s) model (continuous distribution of sedimentation coefficient) and then converted to the c(M) model (continuous distribution of molecular weight). Time invariant noises and baseline offsets were corrected before fitting. A maximum entropy regularization confidence level of 0.95 was used in all the size distribution analyses.

Polarization fluorescence assay. Polarization fluorescence was measured at room temperature with a Synergy 2 (BioTek). All DNA oligonucleotides were purchased from Integrated DNA Technologies without further purification. The three probes have the following sequences: CGCAGTGTAAATTACCTCGAC (Probe-1), CGCAGTGTACTTACCTCGAC (Probe-2), and CAGGCAGCAGGTGTTGGACT (negative control). The 3' terminus of each probe was labeled with fluorescein. The dsDNA samples were prepared by mixing equimolar single-stranded DNA with its complementary sequence. The mixture was heated to 95 °C for 10 min and gradually cooled down to room temperature. dualENH was serially diluted in buffer containing 20 mM Tris-HCl and 100 mM NaCl at pH 8.0, except for the NaCl-dependent experiments in Fig. 4-11A. Concentrations of all probes were kept at 25 nM. The total volume of each sample was kept at 200 µl. The measurements were taken after about a 10-min equilibration. The G factor was calibrated and kept at 0.87 for all samples.

FRET assay. The FRET emission spectrum was measured at room temperature with a Safire2 (Tecan) plate reader. The (TAA)₅ oligonucleotide (oligo) was labeled at its 5' terminus with either Cy3 or Cy5. Preparations of dsDNA samples were made as described above. Samples for the FRET

experiment were prepared by mixing 400 nM Cy3-(TAA)₅ and 600 nM Cy5-(TAA)₅ for a reference, and then by adding 4 μM dualENH to observe a FRET signal change. The buffer contained 20 mM Tris-HCl and 100 mM NaCl at pH 8.0.

Microscope imaging. All imaging was performed at room temperature on a standard epifluorescence microscope (IX71, Olympus) equipped with bright-field and fluorescence modalities. The imaging objective was a 40X NA 0.75 objective lens (UPLFLN 40X, Olympus). The (TAA)₅ oligos were labeled with Cy3 at their 5' terminus. The 11-nt motif oligos (CGCAGTGTAATTTAATTTCTCGAC) were labeled with fluorescein at their 5' terminus. The dsDNA samples were prepared as described above. All experiments were done in the following buffer: 20 mM Tris-HCl, 100 mM NaCl at pH 8.0, except that the NaCl concentration was increased to 150 mM in Fig. 4-11B.

Atomic force microscopy. Samples were deposited on a mica surface in a buffer containing 100 mM NaCl, 4 mM MgCl₂, and 20 mM TrisHCl at pH 8.0. After a 2-min incubation, the mica surface was washed with 3 ml pure water and air-dried. AFM images were taken using repulsive AC mode on an Asylum MFP-3D-bio imager, with an AFM tip spring constant of 3 N/m. The scanning rate was 1 Hz.

X-ray crystallography. The dsDNA used for crystallization had forward and backward sequences as follow: GTGTAATTTAATTTCC and CGGAAATTAATTACA. An equimolar mixture of the forward and backward oligomers was heated to 95 °C for 10 min and gradually cooled down to room temperature. The dualENH (4.9 mM in 1.6 M sodium chloride and 20 mM MES buffer at pH 5.8) and the dsDNA (5.6 mM in 10 mM TrisHCl at pH 8.0) were mixed in equal volumes. Protein-DNA co-crystals were grown at room temperature in 0.2 M potassium thiocyanate and 20% w/v polyethylene glycerol 3350 at pH 7.0 using hanging-drop diffusion. Diamond-like crystals appeared within 1-2 days. The crystals were soaked in 25% ethylene glycol cryoprotectant and flash frozen by liquid nitrogen. Diffraction data were collected at beamline 12-2 at Stanford Synchrotron

Radiation Lightsource. The best diffraction data had a resolution of ~ 3.2 Å. Phases were obtained through molecular replacement using wild-type ENH-DNA co-crystal structure (PDB code: 3HDD) (28) as the searching model. Further refinement was done with PHENIX (36) and Coot (37). The data statistics are listed in Table 4-2.

References

1. Sarikaya M, Tamerler C, Jen AKY, Schulten K, Baneyx F (2003) Molecular biomimetics: nanotechnology through biology. *Nat Mater* 2(9):577-585.
2. Rothmund PWK (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440(7082):297-302.
3. Jiang Q, *et al.* (2012) DNA origami as a carrier for circumvention of drug resistance. *J Am Chem Soc* 134(32):13396-13403.
4. Qian L, Winfree E (2011) Scaling up digital circuit computation with DNA strand displacement cascades. *Science* 332(6034):1196-1201.
5. Gradisar H, *et al.* (2013) Design of a single-chain polypeptide tetrahedron assembled from coiled-coil segments. *Nat Chem Biol* 9(6):362-366.
6. King NP, *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336(6085):1171-1174.
7. Sano T, Smith CL, Cantor CR (1992) Immuno-PCR - very sensitive antigen-detection by means of specific antibody-DNA conjugates. *Science* 258(5079):120-122.
8. Niemeyer CM, Koehler J, Wuerdemann C (2002) DNA-directed assembly of bienzymic complexes from in vivo biotinylated NAD(P)H : FMN oxidoreductase and luciferase. *Chembiochem* 3(2-3):242-245.
9. Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B (2011) Computational design of a symmetric homodimer using β -strand assembly. *Proc Natl Acad Sci USA* 108(51):20562-20567.

10. Fleishman SJ, *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332(6031):816-821.
11. Ashworth J, *et al.* (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* 441(7093):656-659.
12. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struc Biol* 14(2):208-216.
13. Röthlisberger D, *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453(7192):190-195.
14. Privett HK, *et al.* (2012) Iterative approach to computational enzyme design. *Proc Natl Acad Sci USA* 109(10):3790-3795.
15. Jiang L, *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319(5868):1387-1391.
16. Siegel JB, *et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329(5989):309-313.
17. Tinberg CE, *et al.* (2013) Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* 501(7466):212-216.
18. Brennan RG, Matthews BW (1989) The Helix-Turn-Helix DNA-Binding Motif. *J Biol Chem* 264(4):1903-1906.
19. Guerrero L, Smart OS, Woolley GA, Allemann RK (2005) Photocontrol of DNA binding specificity of a miniature engrailed homeodomain. *J Am Chem Soc* 127(44):15624-15629.
20. Marshall SA, Morgan CS, Mayo SL (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol* 316(1):189-199.
21. Shah PS, *et al.* (2007) Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol* 372(1):1-6.
22. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO (1994) Structural studies of the engrailed homeodomain. *Protein Sci* 3(10):1779-1787.

23. Huang P-S, Love JJ, Mayo SL (2005) Adaptation of a fast Fourier transform-based docking algorithm for protein design. *J. Comput. Chem.* 26(12):1222-1232.
24. Mou Y, Mayo SL (Using molecular dynamics simulations to predict domain swapping of computationally designed protein variants. *submitted to J. Mol. Biol.*
25. Mou Y, Huang P-S, Hsu F-C, Huang S-J, Mayo SL (Computational design and experimental verification of a symmetric protein homodimer. *submitted to Proc. Natl Acad. Sci. USA.*
26. Kalionis B, Ofarrell PH (1993) A Universal Target Sequence Is Bound in-Vitro by Diverse Homeodomains. *Mech Develop* 43(1):57-70.
27. King NP, *et al.* (2014) Accurate design of co-assembling multi-component protein nanomaterials. *Nature* 510(7503):103-108.
28. Fraenkel E, Rould MA, Chambers KA, Pabo CO (1998) Engrailed homeodomain-DNA complex at 2.2 angstrom resolution: A detailed view of the interface and comparison with other engrailed structures. *J Mol Biol* 284(2):351-361.
29. Whaley SR, English DS, Hu EL, Barbara PF, Belcher AM (2000) Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly. *Nature* 405(6787):665-668.
30. Ringler P, Schulz GE (2003) Self-assembly of proteins into designed networks. *Science* 302(5642):106-109.
31. Li M, Mann S (2004) DNA-directed assembly of multifunctional nanoparticle networks using metallic and bioinorganic building blocks. *J Mater Chem* 14(14):2260-2263.
32. Li M, Wong KKW, Mann S (1999) Organization of inorganic nanoparticles using biotin-streptavidin connectors. *Chem Mater* 11(1):23-26.
33. Allen BD, Mayo SL (2006) Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* 27(10):1071-1075.

34. Dunbrack RL, Jr., Karplus M (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230(2):543-574.
35. Stemmer WP, Cramer A, Ha KD, Brennan TM, Heyneker HL (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164(1):49-53.
36. Adams PD, *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* 66:213-221.
37. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D* 60:2126-2132.

Table 4-1. Sequences of wild-type ENH and dualENH.



	o ooo	vv vvv vvv	v vv vv	oo oo oo ooo o
ENH (WT)	EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST			
dualENH	-----E--KKA-DLA-YFD----PEW-RY--QR-----			

The three “coils” at the top show the location of the three helices in the wild-type (WT) fold based on PDB structure 1ENH. The DNA-binding domain residues are labeled with “o” and the homodimerization domain residues are labeled with “v”.

Table 4-2. Data collection and initial refinement statistics for the protein-DNA co-crystal structure.

Statistics	Value
Data collection	
Space group	P4222
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	90.1, 90.1, 158.9
α , β , γ (°)	90.0, 90.0, 90.0
Resolution (Å)	39–3.1
R _{merge}	0.043
I/ σ	28.4
Completeness, %	99.8
Multiplicity	12.7
Refinement	
Resolution (Å)	36–3.2
Number of reflections	14287
R _{work} /R _{free} (%)	26/32
Number of molecules in asymmetric unit	8
Number of atoms	2903
Macromolecules	2903
Water	0
Ligands	0
B factors (Å ²)	174.5
Macromolecules	174.5
Water	n/a
Ligands	n/a
R.m.s. deviations	
Bond lengths (Å)	0.012
Bond angles (°)	1.32

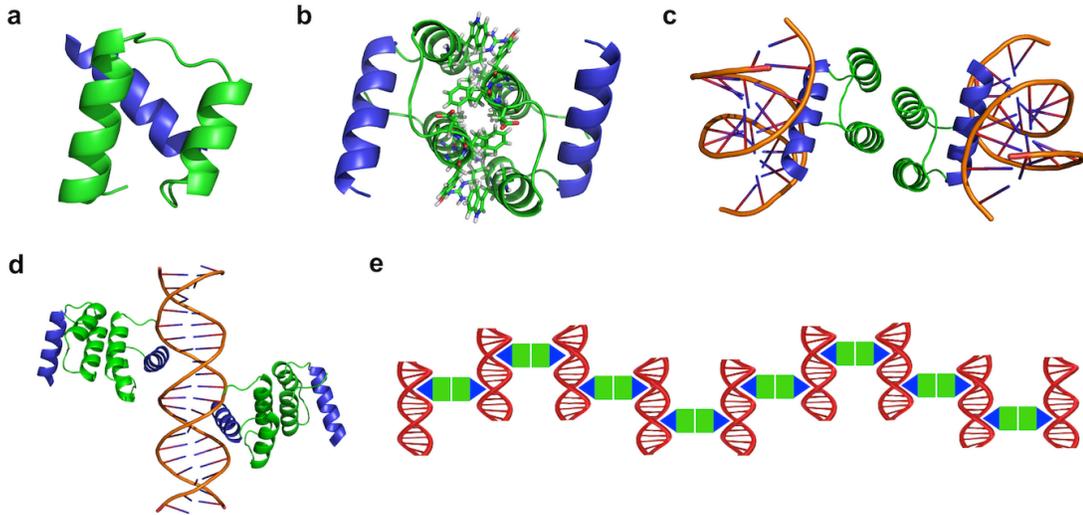


Fig. 4-1. Protein-DNA nanomaterial design strategy. (A) Helix-1 and helix-2 (green) of ENH were engineered into a homodimerization domain, and helix-3 (blue) is the native DNA-binding domain. (B) The interface of the docked model was designed for homodimerization. (C) The designed homodimer, named dualENH, binds two dsDNA fragments on its outward faces. This model was generated by aligning the homodimer model in **b** with the ENH-DNA co-crystal structure (PDB code: 3HDD). (D) Two protein binding sites were engineered onto a dsDNA fragment so that two dualENH dimers would bind 180° apart along the double-helix. (E) The dualENH protein in **c** and the dsDNA fragment in (D) co-assemble into a protein-DNA nanowire. Note that this 2D cartoon is for purposes of illustration only, and that the 3D design model of the nanowire is spiraled.

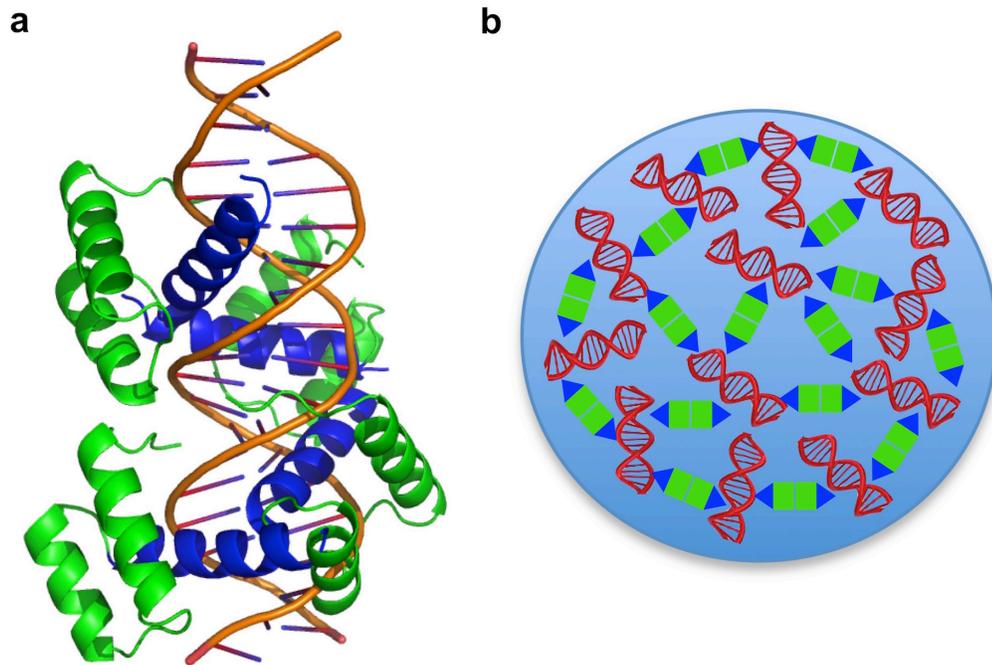


Fig. 4-2. Design model of irregular bulk protein-DNA nanoparticle. (A) Four consecutive ENH binding sites that each face in a different direction are engineered onto a dsDNA fragment. This dsDNA building block allows the protein-DNA assembly to occur in all three dimensions. Note that in this particular design, two neighboring binding sites may not be simultaneously occupied due to steric hindrance. (B) Cartoon illustrating an irregular shaped nanoparticle formed by co-assembly of dualENH and the dsDNA shown in (A) The DNA-binding domains of dualENH are shown as blue triangles, and the homodimerization domains are shown as green squares.

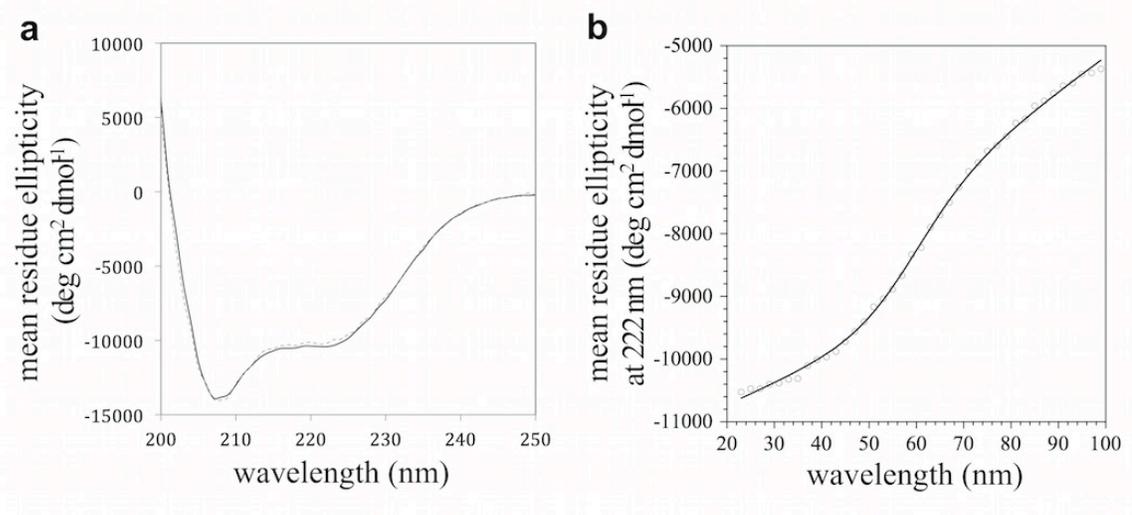


Fig. 4-3. Circular dichroism (CD) spectroscopy of dualENH. (A) CD spectrum of dualENH at room temperature. Solid line: before thermal denaturation; dashed line: after thermal denaturation. The overlapping of the two curves indicates that dualENH folds reversibly. (B) Thermal denaturation curve measured at 222 nm. Circles: experimental data; line: curve fit with two-state transition model. The melting temperature of dualENH was determined to be 59 °C.

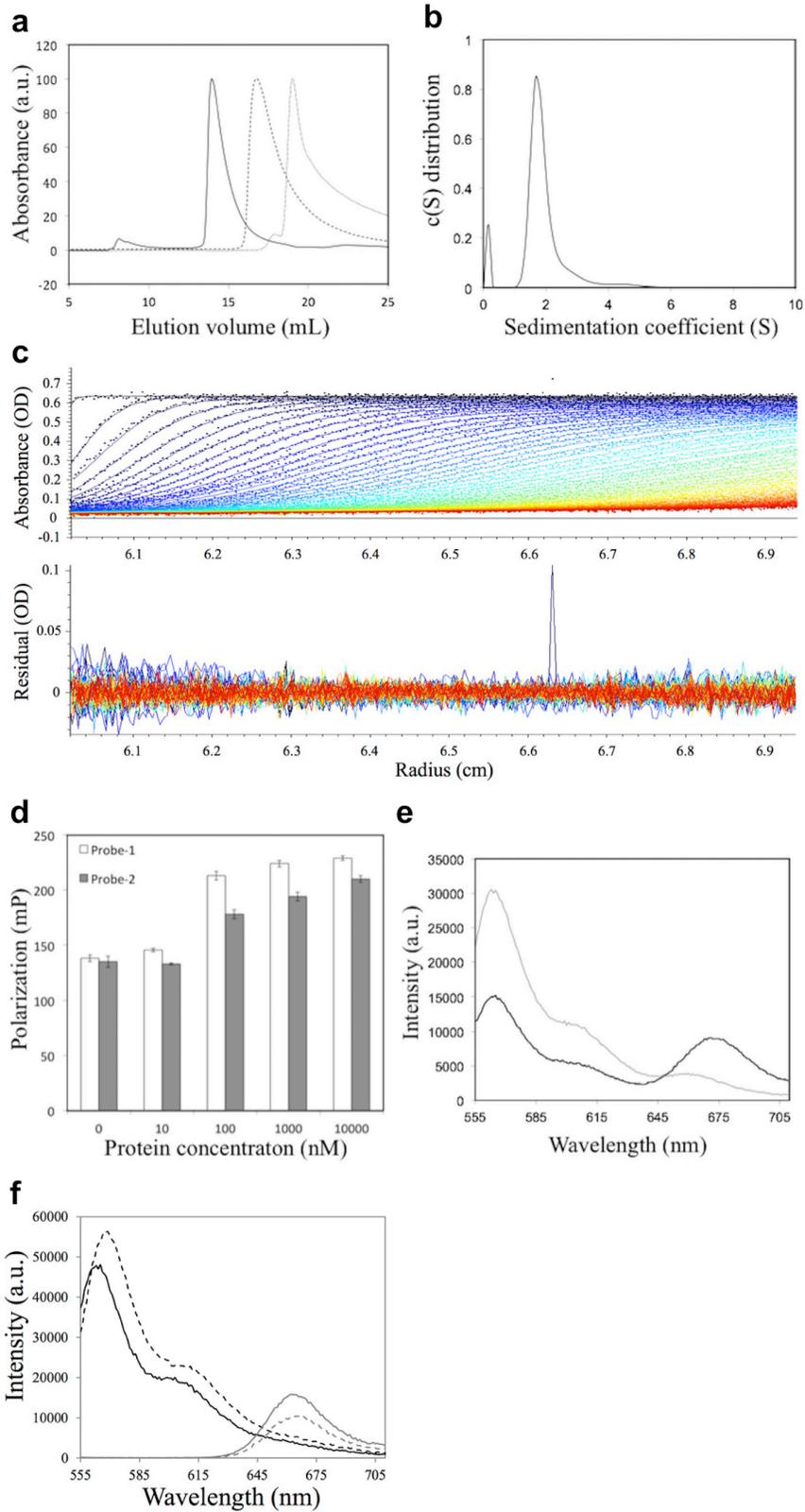


Fig. 4-4. Biophysical characterization of dualENH. (A) Size-exclusion chromatography of dualENH with three different loading concentrations: solid line: 650 μM ; dashed line: 80 μM ; dotted line: 5 μM . The highest signals were normalized to 100 for all curves. (B) $c(S)$ model fit from a sedimentation velocity experiment of 40 μM dualENH. The major peak around $S = 1.9$ corresponds to a MW of 18.3 kD, which is about twice that of monomeric dualENH (8.7 kD). The spike at the left ($S < 0.5$) may be due to impurities or artifacts from model fitting. (C) Raw data and fitting residuals for the sedimentation velocity experiment in (B). A total of 378 curves were used for fitting, but for visual clarity only 1/5th of the curves are shown. The upper graph shows the raw data (dots) and the fitting curves; the lower figure shows the residuals between the experimental data and the fit. The square root of variance of the fit is 0.00669. (D) Fluorescence polarization experiment. Two dsDNA sequences labeled with fluorescein were used as probes to assay dualENH-DNA binding. Probe-1: 20-nt dsDNA with the binding motif TAATTA; probe-2: same sequence as probe-1 but with a single-nucleotide mutation to the binding motif (TA[C]TTA). The concentration of dualENH was varied, while the concentration of the three probes remained constant (25 nM). Data are shown as mean \pm s.e.m. for 3 replicates (E) FRET experiment showing that dualENH brings two dsDNA fragments within Förster distance. 15-nt dsDNA (TAA)₅ were labeled with either Cy3 or Cy5 to serve as the FRET donor or acceptor. Gray line: 400 nM Cy3-(TAA)₅ + 600 nM Cy5-(TAA)₅; black line: 400 nM Cy3-(TAA)₅ + 600 nM Cy5-(TAA)₅ + 4 μM dualENH. (F) Two control experiments for the FRET experiment in e. Black line: 400 nM Cy3-(TAA)₅; Black dashed line: 400 nM Cy3-(TAA)₅ + 4 μM dualENH; Gray line: 600 nM Cy5-(TAA)₅; Gray dashed line: 600 nM Cy5-(TAA)₅ + 4 μM dualENH.

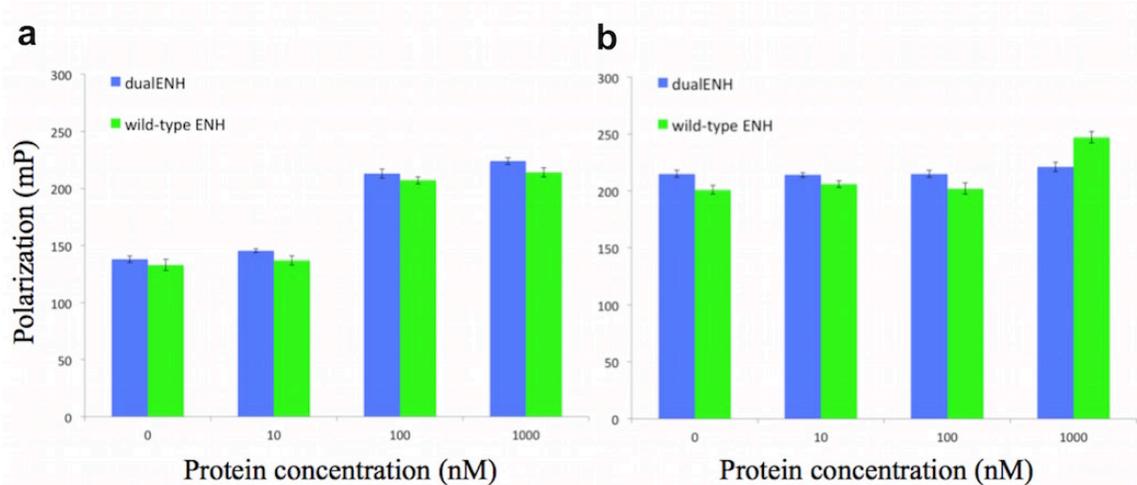


Fig. 4-5. Fluorescence polarization experiments with dualENH and wild-type ENH. (A) Fluorescence polarization experiment with probe-1, which contains the binding motif TAATTA of ENH. The concentration of dualENH or wild-type ENH was varied, while the concentration of probe-1 remained constant (25 nM). Both dualENH and wild-type ENH show saturated binding when the protein concentration is at or above 100 nM. Data are shown as mean \pm s.e.m. for 3 replicates. (B) Same experiments as (A), except that a probe without any TAATNN binding motif was used. Note that probe-1 used in (A) has a lower fluorescence intensity and polarization than the probe used in (B) likely due to partial quenching by a guanine nucleotide on the strand opposite the fluorescein label. Data are shown as mean \pm s.e.m. for 3 replicates.

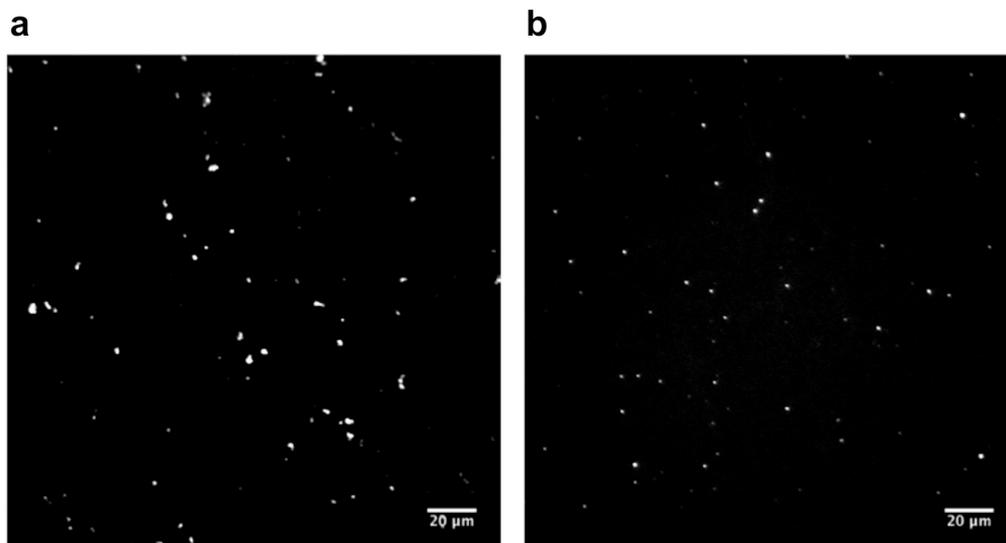


Fig. 4-6. Fluorescence microscopy of protein-DNA nanoobjects. (A) dsDNA (TAA)₅ fragments were labeled with the fluorescent dye Cy3. A fluorescent image was taken of particles formed by mixing 5 μM dualENH with 2 μM Cy3-(TAA)₅ in 20 mM Tris-HCl and 100 mM NaCl at pH 8.0. The particles formed irregular shapes up to ~5 μm in diameter. (B) Same experiment as in (A) except that 25-nt dsDNA fragments containing motif-11 (TAATTTAATTT) in the middle (CGCAGTGTAATTTAATTTCCTCGAC) were used instead of (TAA)₅ fragments. All particle sizes are under the diffraction limit (submicron). The shapes of the particles are slightly oval instead of being symmetrical (circular) due to moderate geometrical aberrations of the microscopy system.

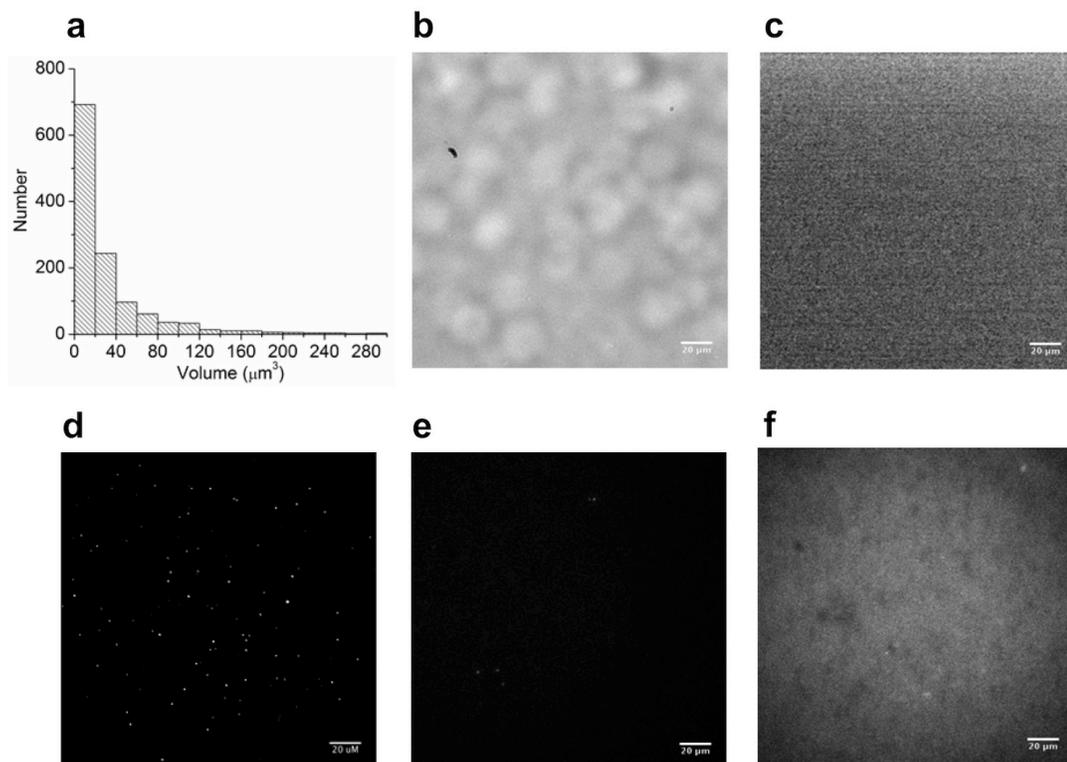


Fig. 4-7. Microscope imaging experiments. (A) The size distribution of the irregular protein-DNA particles formed by 5 μM dualENH mixed with 2 μM Cy3-(TAA)₅. (B) Bright-field microscope image of 5 μM dualENH mixed with 2 μM Cy3-(TAA)₅. A dust particle (upper-left) is evident, indicating the focal plane is correct. (C) Fluorescence microscope image of 2 μM Cy3-(TAA)₅ alone. (D) Fluorescence microscope image of particles formed with 500 nM dualENH mixed with 200 nM Cy3-(TAA)₅. (E) Fluorescence microscope image of particles formed with 200 nM dualENH mixed with 100 nM Cy3-(TAA)₅. (F) Fluorescence microscope image of particle inhibition experiments. A small amount (5 nM) of single-binding-site dsDNA (containing only one TAATTA motif) was pre-mixed with 500 nM dualENH, then 200 nM Cy3-(TAA)₅ was added. The illumination/camera sensitivity was enhanced to confirm that particle formation is nearly completely absent under these conditions.

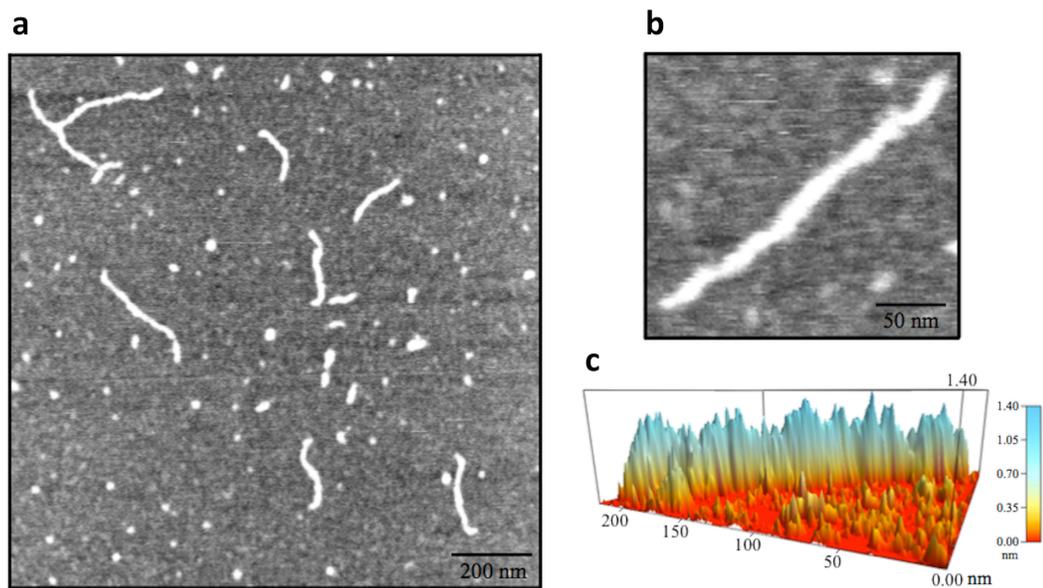


Fig. 4-8. Atomic force microscopy of protein-DNA nanowires. (A) Representative AFM image obtained after mixing 5 μM dualENH with 2 μM of the two-binding-site dsDNA (25-nt dsDNA containing motif-11). Nanowire structures ~ 15 nm wide and up to 300 nm long are clearly visible. (B) Magnified image of a single nanowire ~ 250 nm in length. (C) 3D topology display of (B) shows the height of the nanowire is ~ 1.0 nm.

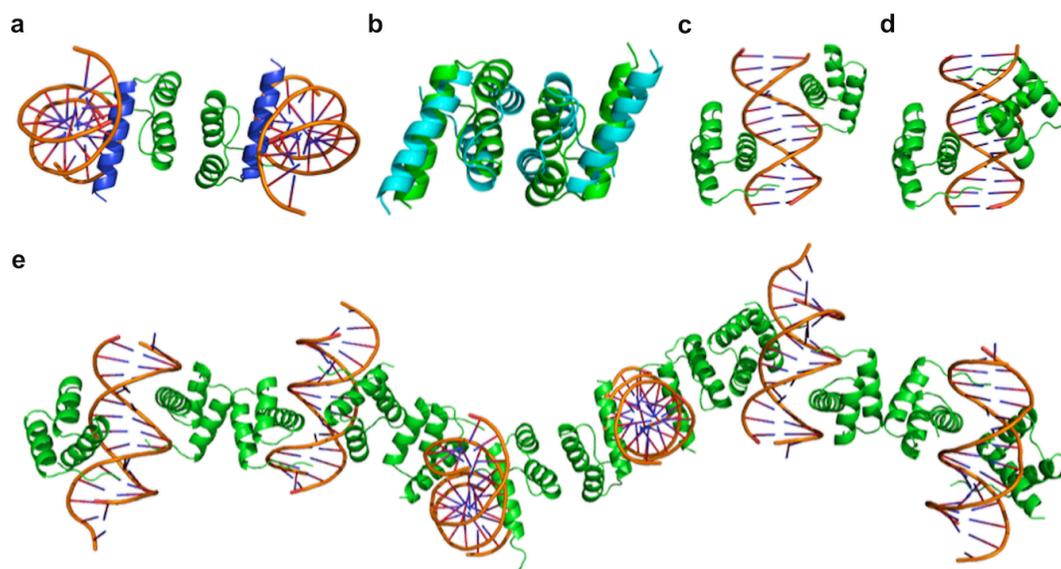


Fig. 4-9. Co-crystal structure of protein-DNA complex. (A) dualENH forms a symmetric homodimer using helix-1 and helix-2 (green) as the protein-protein interface. Helix-3 (blue) binds to the dsDNA in the same way that wild-type ENH does. (B) Two forms of dualENH are present in the co-crystal structure and occur in a molar ratio of 3:1 (green:cyan). (C), (D) Two forms of protein-DNA binding are observed in the co-crystal structure and occur in a molar ratio of 1:1. Both have two dualENH homodimers bound on the designed 11-nt motif TAATTTAATTT. In (C), both of the dualENH dimers bind in the optimal motif (TAATTT) orientation, whereas in (D), one of the dualENH dimers (right) binds in the suboptimal orientation (AAATTA, the reverse complementary sequence of the optimal motif). (E), Slightly kinked nanowire structure found in the co-crystal structure.

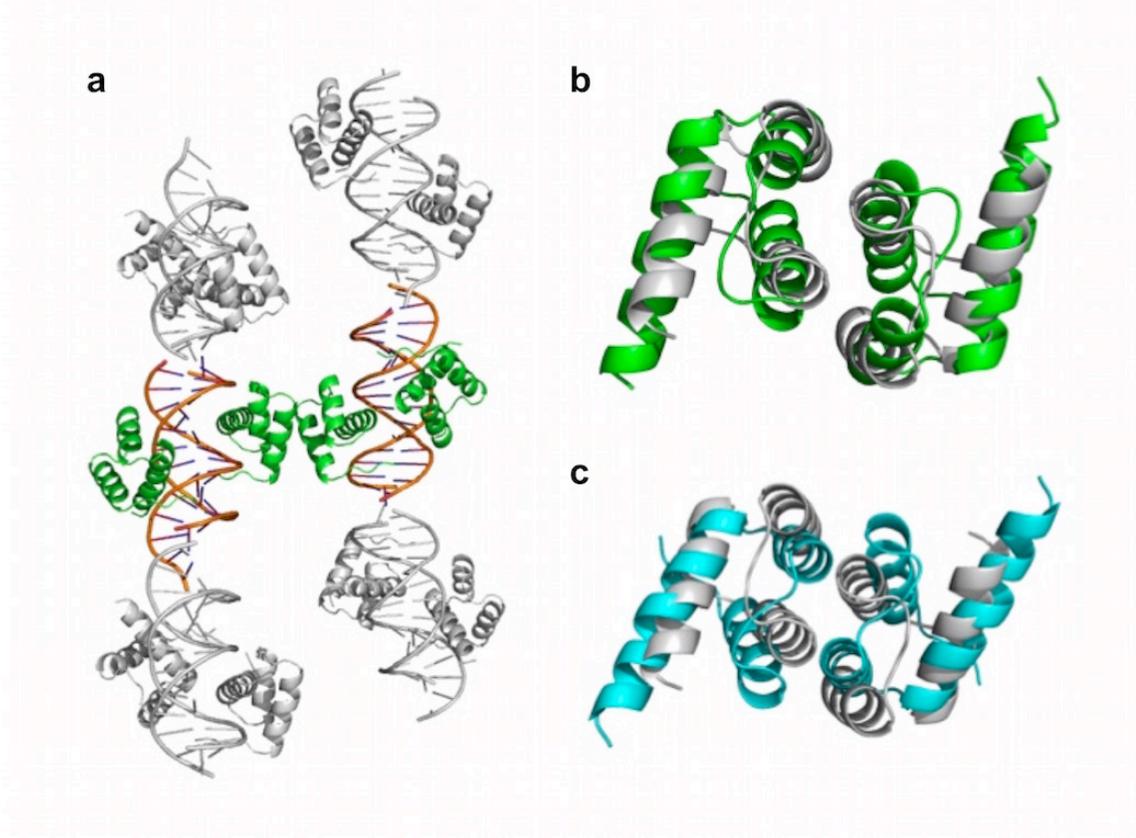


Fig. 4-10. Co-crystal structure of the protein-DNA complex. (A) structures in the asymmetric unit cell are shown in color, and the end-to-end packing of neighboring DNA molecules and their bound proteins are shown in gray. (B), (C) the dualENH homodimer observed in the co-crystal structure (green or cyan) is superimposed with the design model (gray). The backbone RMSD to the design model is 3.8 Å (green) and 3.9 Å (cyan), respectively. When only one subunit is aligned between the more predominant configuration (green) and the design model, the angular displacement between the other subunits is $\sim 45^\circ$ with about 3 Å translational displacement. The less predominant configuration has a lower angular displacement $\sim 20^\circ$ but a larger translational displacement ~ 8 Å. The calculated energies for the design model and the two crystallographic dimers are -155.2, -140.3 (green) and -131.2 (cyan) Rosetta energy units, respectively.

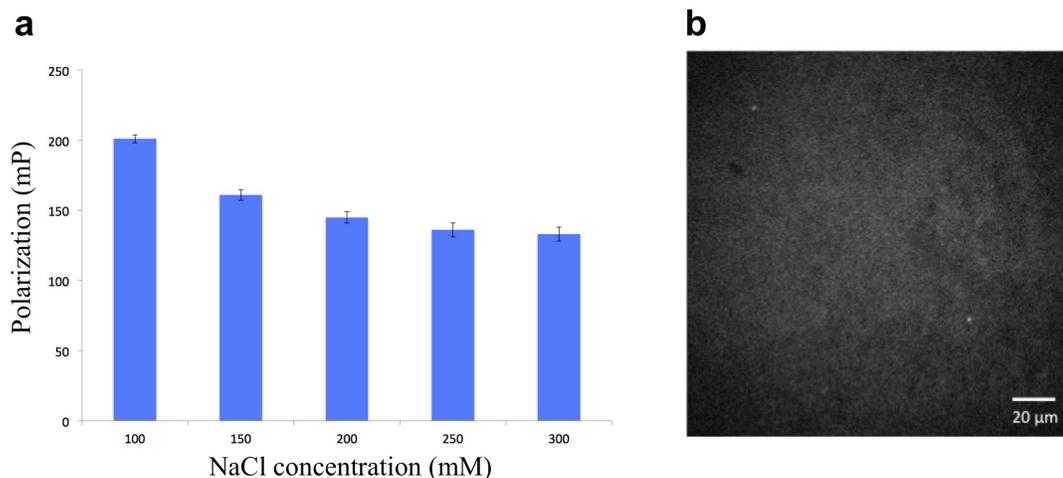


Fig. 4-11. dualENH-DNA binding and nanostructure formation are inhibited at high salt concentrations. (A) Fluorescence polarization experiments of dualENH and probe-1 at various NaCl concentrations. Probe-1 and dualENH were mixed in buffers with different NaCl concentrations and fluorescence polarization values were recorded. DualENH-DNA binding dropped significantly from 100 mM to 150 mM NaCl, and was completely absent at 300 mM NaCl. Data are shown as mean \pm s.e.m. for 3 replicates. (B) Fluorescence image of particle experiment at 150 mM salt concentration. The sample was prepared by mixing 500 nM dualENH and 200 nM Cy3-(TAA)₅ in 20 mM Tris-HCl buffer with 150 mM NaCl. The illumination/camera sensitivity was enhanced to confirm that particle formation is nearly completely absent under these conditions.

Appendix

Direct Visualization Reveals Dynamics of a Transient Intermediate During Protein Assembly

The text of this chapter was adapted from a manuscript coauthored with Xin Zhang, Vinh Q. Lam, Tetsunari Kimura, Jaeyoon Chung, Sowmya Chandrasekar, Jay R. Winkler, Stephen L. Mayo, and Shu-ou Shan

Xin Zhang, Vinh Q. Lam^{*}, Yun Mou^{*}, Tetsunari Kimura, Jaeyoon Chung, Sowmya Chandrasekar, Jay R. Winkler, Stephen L. Mayo, and Shu-ou Shan (2011) Direct visualization reveals dynamics of a transient intermediate during protein assembly. *Proceedings of the National Academy of Sciences USA* 108, 6450-6455. (* these authors contributed equally)

Reproduced with permission.

Abstract

Interactions between proteins underlie numerous biological functions. Theoretical work suggests that protein interactions initiate with formation of transient intermediates that subsequently relax to specific, stable complexes. However, the nature and roles of these transient intermediates has remained elusive. Here, we characterized the global structure, dynamics, and stability of a transient, on-pathway intermediate during complex assembly between the Signal Recognition Particle (SRP) and its receptor (SR). We show that this intermediate has overlapping but distinct interaction interfaces from that of the final complex, and is stabilized by long-range electrostatic interactions. A wide distribution of conformations is explored by the intermediate; this distribution becomes more restricted in the final complex and is further regulated by the cargo of SRP. These results suggest a funnel-shaped energy landscape for protein interactions, and provide a framework for understanding the role of transient intermediates in protein assembly and biological regulation.

Introduction

Interactions between proteins are central to biology, and underlie numerous molecular recognition, regulation, and signaling events (1). A challenge in our understanding of protein interactions is to reconcile their fast association kinetics required for biological function (10^6 – 10^8 $M^{-1} s^{-1}$) with the fact that formation of stable protein assemblies often involves extensive short-range, stereospecific interactions that are difficult to accomplish during a single diffusional encounter (2-4). This problem becomes more pronounced in protein interactions that require extensive conformational changes in the interaction partners. Much theoretical work has suggested that assembly of a protein complex initiates with the formation of a transient intermediate held together by solvent cage and long-range electrostatic attractions, followed by relative rotatory diffusions of the binding partners to search for the optimal interaction interface with shape and electrostatic complementarity (4-9). An extreme example of this concept is the “fly-casting mechanism”, in which unstructured protein molecules bind targets weakly at a relatively large distance, followed by folding at the target site (10-12). In general, formation of transient intermediates reduces the dimension of translational and rotational search and could significantly accelerate protein association.

Despite significant progress in theoretical work, direct experimental demonstration of this model has been limited, and the structural and dynamic nature of transient intermediates during protein interactions has remained elusive. Experimental studies of transient intermediates are still at the infant stage, because these intermediates have short lifetimes and are rarely populated at equilibrium. Pioneering NMR studies have revealed the structures of rare conformational states in equilibrium with the predominant structure in the apo-protein or the final complex, and provided direct experimental support for the ability of proteins to explore different conformations (13-17). Nevertheless, many of these studies have focused on protein interactions that are inherently weak and nonspecific; whether the same principle applies to the assembly of a stable and stereospecific protein complex remains to be determined. Further, the transient species probed in this manner do

not necessarily represent on-pathway intermediates during complex assembly. To understand the protein assembly pathway, it is crucial that on-pathway intermediates during protein assembly can be isolated. To this end, we chose the interaction between the Signal Recognition Particle (SRP) and the SRP receptor (SR) as a model system.

Rapid assembly of a stable SRP•SR complex is required to efficiently deliver cargo proteins to cellular membranes during co-translational protein targeting, and is essential for proper protein localization in all cells (18, 19). Formation of a stable SRP•SR complex is mediated by specific interactions between their NG domains (comprised of a GTPase, G-domain, and a helical N-domain) (Fig. A-1A). However, free SRP and SR are not in the optimal conformation to bind one another, and extensive rearrangements must occur in both proteins to attain a stable complex (20). Previous kinetic studies showed that stable SRP-SR complex assembly begins with the formation of a transient “early” intermediate (Fig. A-1A, step 1), which forms quickly ($k_{on} = 5.8 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$) but is unstable ($K_d \sim 4\text{--}10 \text{ }\mu\text{M}$ and $k_{off} \sim 62 \text{ s}^{-1}$) (21). This intermediate then slowly rearranges ($k_2 \sim 1.5 \text{ s}^{-1}$) to form the final stable complex, which is stabilized by a large, continuous interaction surface between the NG-domains of both proteins (Fig. A-1A, left panel and step 2) (21). Importantly, complex assembly can be stalled at the early intermediate stage by leaving out guanosine 5'-triphosphate (GTP) while maintaining the kinetic competence of this intermediate (Fig. A-1A) (21). This allowed us to isolate this intermediate and directly characterize its global structure, dynamics, and stability in this work. The results identified distinct interaction interfaces used by the early intermediate, and provided direct evidence for extensive conformational search in this intermediate and the importance of long-range electrostatic interactions in its stability. Further, the conformational distribution of the early intermediate is exquisitely sensitive to the biological cues of the SRP, providing potential mechanisms for biological regulation.

Results

Map the binding interface of the early intermediate

The interaction surface in the stable SRP•SR complex is formed primarily by close contacts between their G-domains, with limited contacts between the N-domains near the N-G domain interface contributing the remainder of the interface (Fig. A-1A) (20, 22). We first asked whether the early intermediate forms the same or distinct interaction interface. To this end, electron paramagnetic resonance (EPR) spectroscopy was used to probe the interaction surface (23-25). Based on the co-crystal structure of the stable SRP•SR NG-domain complex, we selected residues in the vicinity of the interaction surface on SR for replacement by cysteine, which allowed site-directed spin labeling with the nitroxide probe (1-oxy-2,2,5,5-tetramethyl-3-pyrrolynyl-3-methyl) methanethiosulfonate (MTSSL). Only the sites where cysteine replacement and nitroxide labeling did not substantially affect the SRP-SR interaction were used for EPR measurements (Fig. A-2A and B). The residues on or near the dimer interface are likely to undergo significant changes in spin probe mobility upon complex formation. These changes are measured by the linewidth of the central resonance (Fig. A-1B and Fig. A-2C and D, DH_0) and the overall breadth of the EPR spectra, especially the intensity of hyperfine splitting that arises from highly immobile populations of spin probes relative to the mobile population (Fig. A-1B, dashed vs. solid arrows) (23).

To validate this approach, we first characterized the interaction surface of the stable complex formed with a non-hydrolyzable GTP analogue, 5'-guanylylimido-diphosphate (GMPPNP). Twelve residues underwent significant EPR spectral changes upon complex formation (Figs. A-1B and A-2C, red vs. black). The majority of spin probes at these positions underwent changes in both central linewidth, and the relative population of immobile species (Fig. A-1B). For two of these residues, T356 and N426, changes in probe mobility was not obvious from the central linewidth but was detectable from changes in EPR spectral shape (Figs. A-1B and A-2C). Collectively, these data provided a view of the interaction surface in the stable

complex that is consistent with the co-crystal structure (Fig. A-1C, black vs. red outline) (20). This validated EPR as a powerful tool to probe the interaction surface of the complex.

We next used this approach to locate the interface of the early intermediate. Three classes of residues were identified that underwent distinct EPR spectral changes upon formation of the early or stable complex (Fig. A-1B). Residues in class I, represented by V242 (Figs. A-1B and A-2C, purple), underwent similar reductions in spin probe mobility upon formation of both the early and stable complexes, suggesting that they are involved in the interface of both complexes. Residues in class II, represented by S429 (Figs. A-1B and A-2C, brown), underwent substantial immobilization of the spin probe in the early intermediate, but these probes became more mobile in the stable complex, suggesting that they are engaged in stronger (in the cases of S429 and T451) or distinct patterns (in the cases of L433 and E487) of interactions in the early intermediate. Residues in class III, represented by I237 (Figs. A-1B and A-2C, green), exhibited substantial changes in spin probe mobility only in the stable complex, suggesting that they are specifically involved in the formation of the stable complex.

Collectively, eight residues underwent substantial spectral changes upon formation of the early intermediate (Fig. A-1B, classes I and II). Compared to the stable complex, these residues reside primarily in or near the N-domain (Fig. A-1C, right panel), suggesting that the early intermediate has a detectable interaction surface that partly overlaps with, but is distinct from that of the stable complex (Fig. A-1C). In addition, some of the residues that change mobility specifically in the stable complex (I237, Q425, N426; Figs. A-1C and A-3, green) were in or adjacent to the conserved 'TAKGG' and 'QLLIADV' motifs, which act as a hinge at the N-G domain interface to readjust the relative orientation of the G- and N-domains during stable complex formation (20). The absence of significant spectral changes at these positions in the early intermediate suggests that this crucial rearrangement has not taken place at the early intermediate stage.

To independently identify the interaction surface of the early intermediate, we introduced 24 mutations in SR, all of which map to the heterodimer interface in the stable complex (Fig. A-4A). These mutations disrupt either the interactions at the dimer interface or the rearrangement at the N-G domain interface, and each impairs formation of the stable complex by 5–200 fold (22). Several of them were also near the residues engaged in the dimer interface of the early intermediate, as identified by EPR (cf. Fig. A-4A vs. Fig. A-1C, right). We tested whether these mutations disrupted the stability of the early intermediate using fluorescence resonance energy transfer (FRET) between coumarin (DACM) labeled SRP C235 and BODIPY-fluorescein (BODIPY-FL) labeled FtsY C487 (21). To our surprise, most of these mutations did not disrupt the early intermediate (Fig. A-4B, black bars). Only three mutations caused moderate reductions in the stability of the early intermediate (2–4 fold), and a combination of all three mutations destabilized the early intermediate by only 8-fold (Fig. A-4B, gray bars and Fig. A-4C).

The mutational analyses provided independent support for important conclusions from the EPR experiments, including the paucity of G-domain interactions and the absence of conformational readjustments at the N-G domain interface in the early intermediate. On the other hand, these data raised additional questions, as they showed that the early intermediate was insensitive to many mutations near its putative interaction surface identified by EPR. Two models to reconcile these results were tested and verified in the experiments below. First, the major interactions that stabilize the early intermediate may lie further outside the G-domain and its vicinity, where most of the mutations above were located. Second, the early intermediate may not have a defined structure but rather contain multiple conformations, each with a distinct interface. Mutations that disrupt a specific interface do not affect alternative conformations, and hence do not significantly affect the overall stability of the intermediate. In contrast, the stable complex has a more defined structure, and hence is more susceptible to mutations that disrupt its interface.

Conformational dynamics in the early intermediate

To test whether the early intermediate samples a broad distribution of conformations, time-resolved FRET (TR-FRET) was used to measure the distance distribution between donor (DACM) and acceptor (BODIPY-FL) dyes labeled at specific sites on SRP and SR in different SRP•SR complexes. These measurements provided nanosecond snapshots of fluorescence decay of the donor dye (Fig. A-5), from which donor-acceptor distance distributions of the respective complex could be derived (26). We analyzed the fluorescence decay curves using both the least-squares fitting (Fig. A-6) and maximum entropy (Fig. A-7) methods. These algorithms produce the narrowest and broadest distance distributions, respectively, that satisfy the experimental measurements, and the distance distributions in the ensemble of SRP-SR complex likely reside in between these two extreme representations. Given this, substantial caution was taken in the interpretation of the distance distributions, such that the conclusions are largely independent of the method used to represent the data. Moreover, we focused on the changes in the distance distribution in the different SRP-SR complexes, which are less sensitive to biases introduced by different data representation.

Three pairs of residues were used to measure distance distributions between the G domains (Fig. A-7A, G-G), the NG domain interfaces (Fig. A-7B, NG-NG), and the N domains (Fig. A-7C, N-N) of both proteins. For all three pairs, the early intermediate exhibited broad distance distributions spanning $\sim 25 - 60 \text{ \AA}$ without a single dominant population (Figs. A-7 and A-6, blue). In contrast, the distributions became significantly more restricted in the stable complex (Figs. A-7 and A-6, red). For the G-G and NG-NG pairs (Figs. A-7A-B and A-6A-B, red), a predominant population was observed in the stable complex with a distance in good agreement with the co-crystal structure (33 \AA in crystal structure vs. 37 \AA for the G-G pair, and 30 \AA in crystal structure vs. 31 \AA for the NG-NG pair) (20). In comparison, the N-N pair displayed a broader distribution in the stable complex, with a peak centered around 37 \AA (Figs. A-7C and A-6C, red). This observation can be partially explained by the limited interactions between the N-

domains in the stable complex, which might allow these domains to have more flexibility (20, 27). Taken together, these results provided direct evidence that the early intermediate contained a large ensemble of conformations that are similar in stability, whereas the stable complex has a more specific structure, particularly at the G-domains and NG-domain interfaces.

Comparison of these distance distributions also provided clues to the complex assembly process. A significant population of molecules with distances as short as ~ 25 Å was observed for the N-N pair in the early intermediate, but this population diminished in the stable complex (Figs. A-7C and A-6C). In contrast, a significant population of molecules exhibited long distances (45–60 Å) for the G-G pair in the early intermediate, which also diminished in the stable complex (Figs. A-7A and A-6A). This suggests that SRP-SR complex assembly initiates from close contacts between their N-domains in the early intermediate, whereas the G-domains are further apart.

Electrostatic interactions drive formation and stability of the early intermediate

Consistent with the notion that complex assembly initiates with contacts between the N-domains, adaptive Poisson-Boltzmann solver (APBS) calculation (28) revealed clusters of positively and negatively charged residues, respectively, on the surface of SRP and SR's N-domains (Fig. A-8). Interactions between these electrostatically complementary surfaces were supported by their evolutionary conservation (Fig. A-9A), and by molecular docking simulations using the ClusPro 2.0 program (29), which generated molecular models for the early intermediate. Two groups, each containing an ensemble of ~ 90 structures, scored significantly higher than all the alternative configurations (representatives in each group are shown in Fig. A-8B). In the 'N' group, the N-domains of SRP and SR contact one another via the electrostatically complementary surfaces identified in the APBS calculation (Fig. A-8B, left); in the 'G' group, the G-domains of the proteins contact one another via an interface that is shifted away from the heterodimer interface in the stable complex (Fig. A-8B, right vs. inset). The nucleotide-binding cavity of SRP and SR

were exposed in both groups (Fig. A-8B, left and right panels), explaining why formation of the early intermediate is a nucleotide-independent process.

Both the ‘N’ and ‘G’ groups represent possible conformations within the ensemble of structures of the early intermediate, as all the residues that changed mobility in EPR measurements resided on the dimer interfaces of one or the other groups (Fig. A-10A and B), and as both groups were needed to reproduce the experimentally observed broad distance distributions of FRET probes (cf. Fig. A-10C-E vs. Fig. A-7, blue). Nevertheless, the following strongly suggest that the ‘N’ group represents the major conformational ensemble. First, the residues that changed spin probe mobility in the early intermediate are primarily in or near the N-domain (Fig. A-1C, right panel). Second, most mutations in the G-domain that could affect the ‘G’ group did not abolish the stability of the early intermediate, suggesting that the conformers in the ‘G’ group are less significantly populated in this intermediate. Third, in Brownian Dynamics calculations (30), the association rate constant for the early intermediate estimated for the ‘N’ group agreed well with the experimental value, whereas that for the ‘G’ group was 30-fold slower (Fig. A-8C).

What features in the ‘N’ group make it the major conformation of the early intermediate? We reasoned that the complementarily charged surfaces on the N-domains of SRP and SR could facilitate the long-range electrostatic interactions that bring the two proteins into proximity (Fig. A-8A). To test the contribution of these electrostatic interactions, we generated charge reversal mutants in which three basic residues (R35, R49 and K56) on the SRP N-domain were mutated to glutamates (RK3E), and the glutamate residues in the EELEE motif on the SR N-domain were mutated to arginines (RRLRR). Mutants SRP (RK3E) and SR (RRLRR) severely reduced the stability of the early intermediate (Figs. A-8D and A-9C); these mutants also caused 10–28 fold reductions in the association rate constant for the stable SRP-SR complex assembly that correlated well with their reduced stabilities in the early intermediate (Fig. A-8D, blue, and Fig. A-9D).

We further asked whether the SRP-SR interaction can be rescued by combining the charge reversal mutants of SRP and SR, which partially restores the electrostatic interaction between their N-domains (Fig. A-9B). Indeed, the combination of the SRP (RK3E) and SR (RRLRR) mutants restored the stability of the early intermediate to within three-fold of that of the wildtype protein (Fig. A-8D, black bars and Fig. A-9C). The kinetics of stable complex assembly was correspondingly rescued (Fig. A-8D, blue bars and Fig. A-9D). The incomplete rescue could be accounted for by the fact that, although the SRP (RK3E) mutation made the SRP N-domain highly negatively charged, the SR (RRLRR) mutation rendered the SR N-domain only moderately positively charged (Fig. A-9B). Together, these results strongly support the notion that electrostatic interactions provide an important driving force to form and stabilize the early intermediate, which correspondingly enhances the kinetics of stable complex assembly.

Cargo restricts conformational dynamics of the early intermediate

The SRP-SR interaction is profoundly influenced by the cargos of SRP, the ribosome-nascent chain complexes (RNCs), which stabilize the early intermediate over 50-fold and accelerate stable complex assembly over 100-fold (18, 19). We speculated that the cargo could actively regulate the conformational dynamics of the early intermediate. To test this hypothesis, we used TR-FRET to measure the conformational distribution of the early intermediate in the presence of RNC_{FtsQ}, which contains the first 74 amino acids of a known SRP substrate FtsQ. Notably, the cargo substantially altered the distance distribution of all the FRET pairs in the early intermediate, changing their broad distance distributions to more bi-modal patterns (Figs. A-7 and A-6, green). Thus the cargo restricts the dynamics of the early intermediate to a more limited conformational space, in which the successful selection of complementary structures might be enhanced. This could partly explain how the cargo enhances the kinetics of SRP-SR complex assembly, and therefore effect efficient protein targeting (18, 19).

Discussion

Using the SRP-SR interaction as a model system, we analyzed the global structure, dynamics, and stability of an on-pathway intermediate during the assembly of a stable protein complex. The techniques used here would not provide atomic resolution information for the assembly intermediate; on the other hand, a combination of biochemical, biophysical, and theoretical approaches provided a set of complementary and self-consistent information that together revealed important global features of this intermediate and shed light on the association process of a relatively large and stable protein complex.

An intriguing finding of this work is that the interaction surface used by the early intermediate is quite distinct from that of the stable complex. Electrostatic interactions between complementarily charged surfaces on the N-domains of SRP and SR provided the primary stabilizing force for the early intermediate. In contrast, more stereospecific interactions between the G-domains, which provide most of the driving force for the stable complex, are rather weak at this stage of assembly. This explains the previous observations that formation of the early intermediate can occur independently of GTP and nucleotides can rapidly exchange in this intermediate (21), and is also consistent with a recent cryo-EM analysis of an early cargo-SRP-SR targeting complex (31). The early intermediate studied herein is considerably more stable ($K_d \sim 4 - 10 \mu\text{M}$ and $k_{off} \sim 62 \text{ s}^{-1}$) than would be expected for encounter complexes, and likely occurs at a stage later than simple diffusional encounter. The fact that this intermediate still has a distinct interaction surface than the final complex strongly suggests that productive protein-protein interactions can initiate at sites that are adjacent to but quite distinct from the final interaction surface.

Besides the electrostatic interaction between the N-domains, a conserved electrostatic interaction between Lys399 in the G-domain of SR and the GGAA tetraloop of the SRP RNA (the other component of SRP) also provides a crucial contact that stabilizes the early intermediate (~ 12 -fold) and accelerates stable complex assembly (32). Despite extensive mutagenesis, these

two pairs of electrostatic interactions are the only ones that have been found thus far to contribute significantly to the stability of the early intermediate. Together, these results show that formation of the SRP-SR early intermediate is driven primarily by long-range electrostatic attractions. Consistent with this notion is that the stability of the early intermediate and the rate of stable complex assembly have a strong dependence on ionic strength (32). Critical roles of electrostatic interactions in enhancing protein interaction kinetics have been predicted theoretically (33, 34) and demonstrated in multiple cases (35-38); our results further emphasize the role of such interactions in stabilizing assembly intermediates, which provides an effective way to accelerate the overall assembly process.

Another intriguing finding here is that TR-FRET measurements revealed a broad conformational distribution for the early intermediate (Fig. A-11, blue). The broad conformational distribution is also supported by the observation that single mutations in the G domain do not significantly affect the stability of the early intermediate, whereas a combination of these mutations causes a substantial disruption of its stability. This provides direct evidence that a wide conformational space is explored by this intermediate, which could aid in the search and selection for the optimal structure conducive to forming the stable complex (39). Interestingly, the conformational space of the intermediate is actively regulated by the cargo of SRP (Fig. 5, green), which restricts the conformation of the early intermediate and produces a more bi-modal pattern of distribution. These changes could potentially provide a mechanism to exert biological regulation (18, 19). Nevertheless, much more work will be needed to provide a molecular understanding of the conformational changes brought upon by the cargo and how these changes affect the complex assembly process.

Formation of the stable complex significantly restricts the distance distributions of both the G-G and NG-NG FRET pairs, consistent with the notion that a stable and stereospecific complex ($K_d \sim 16 - 30$ nM) has a much more defined structure. In comparison, a broader distribution of FRET distances was exhibited by the N-N FRET probes, which might arise from a

combination of the following factors. First, residual, albeit more restricted, conformational sampling still occurs in the stable complex (Fig. A-11, red). Second, interactions in the stable complex primarily involve the G-domains and the NG-domain interface, whereas contacts between the N-domains are rather limited (Fig. A-1A). Thus the N-domains are likely to have more flexibility than the G-domains and can sample different configurations in the stable complex.

The features of the early intermediate during protein assembly bear intriguing analogies to molten globules during protein folding, in that both are relatively resistant to many mutations and have a broad free energy landscape that allows the protein(s) to sample multiple configurations (40). Also analogous to the protein folding process, the energy landscape of protein assembly appears to be funnel-shaped, and becomes narrower as the free proteins transition through the intermediate and progress towards the stereospecific complex (Fig. A-11), as predicted by theoretical work (7, 10-12, 33, 41). These findings could represent general features of transient intermediates during assembly of stable protein complexes, and provide a framework to understand their roles in enhancing protein interactions and biological regulation.

Methods

Materials. The *E. coli* SRP and SR GTPases (Ffh and FtsY, respectively) and the 4.5S RNA were expressed and purified as previously described (21, 42). All the experiments in this work use SRP, which is the complex of Ffh bound to the 4.5S RNA. Truncated FtsY (47-497) was used in all the fluorescence and EPR measurements, except for the charge reversal mutants FtsY (RRLRR). The abilities of FtsY (47-497) to interact with SRP and respond to the cargo are similar to those of full length FtsY (42). Mutant Ffh and FtsY's were constructed using the QuickChange mutagenesis procedure (Stratagene). All the mutant proteins were expressed and purified using the same procedure as that for the wild-type proteins. Fluorescent dyes N-(7-dimethylamino-4-

methylcoumarin-3-yl)maleimide (DACM) and BODIPY-FL-N-(2-aminoethyl)-maleimide were from Invitrogen.

RNC_{FtsQ} purification. Homogeneous RNC_{FtsQ} were generated from *in vitro* translation using membrane free cell extract prepared from MRE600 cells, and purified by affinity chromatography and sucrose gradient centrifugation as previously described (18, 43). Purified RNC_{FtsQ} serves as a functional cargo in protein targeting, as it can bind SRP, trigger factor, and the secYEG translocon complex (43). In quantitative assays, purified RNC_{FtsQ} exhibited the same affinity for SRP as those measured with RNCs that do not contain an affinity tag (44).

Fluorescence labeling. For FRET measurements, DACM and BODIPY-FL were used to label single-cysteine mutants of Ffh and FtsY, respectively, as previously described (21). Labeled protein was purified as described (21), and the efficiency of labeling was typically $\geq 95\%$ with a background of $< 5\%$.

Spin labeling. Single cysteine mutants of FtsY [in 20 mM HEPES (pH 8.0), 150 mM NaCl, and 2 mM EDTA] were incubated with a 10-fold molar excess of dithiothreitol (DTT) at room temperature for 1-2 h to reduce any disulfide bonds. DTT was removed by gel filtration chromatography. The reduced and degassed proteins (~ 100 μ M) were labeled with a 3–5 fold molar excess of MTSSL (Toronto Research Chemicals, Toronto, Canada) at room temperature in the dark for 2-3 h. Excess MTSSL was removed by gel filtration chromatography. The labeling efficiency was determined by EPR using the TEMPO calibration curve (Bruker user manual), and was typically $> 80\%$ with $< 5\%$ background as assessed from the cysteine-less wild-type protein using the same procedure. All the spin-labeled proteins were tested for interaction with SRP using the GTPase assay; only the spin-labeled FtsY mutants that did not substantially disrupt activity were used for EPR measurements.

TR-FRET measurements. Time-resolved fluorescence decay measurements were carried out in SRP buffer with a picosecond streak camera (C5680; Hamamatsu Photonics) in the photon-

counting mode (45), using an excitation wavelength of 355 nm generated from a third harmonic of a regeneratively amplified mode-locked Nd-YAG laser (pulsewidth is ~ 15 ps) (Vangurd, Spectra-Physics). A band-pass filter of 450 ± 5 nm was used as the emission filter. There was no observable fluorescence from buffer or unlabeled protein. DACM fluorescence decay kinetics was measured in both short (5 ns) and long (20 ns) time scale, with time resolutions of ~ 10 and ~ 40 ps, respectively.

TR-FRET conditions. Donor-only measurements were carried out in SRP buffer in the presence of 5 or 1 μM DACM-labeled SRP for the early and stable complexes, respectively. For the early intermediate, 5 μM DACM-labeled SRP and 50 μM BODIPY-FL-labeled SR were mixed together in the presence of GDP. For the stable complex, 1 μM DACM-labeled SRP and 8 μM BODIPY-FL-labeled SR were mixed in the presence of GMPPNP. Under these conditions, formation of both complexes was complete after a 20-minute incubation at room temperature in dark.

Numerical analysis for TR-FRET measurements. The measured short and long time-scale data were spliced together, and the combined traces were compressed logarithmically before the fitting process (70 points per decade). The splicing and compression did not introduce artifacts to the interpretation of data (26). Analyses of the TR-FRET data can be described as a numerical inversion of a Laplace transform $[I(t) = \sum_k P(k) \exp^{-kt}]$, in which $I(t)$ is fluorescence intensity, k is the fluorescence decay rate constant, and $P(k)$ is the probability of a specific k (46, 47). In this work, two algorithms were used to invert the kinetics data with regularization methods that also impose a non-negativity constraint, $P(k) \geq 0$ ($\forall k$). The first method, based on the Least-Squares (LSQ) fitting, used a MATLAB algorithm (LSQNONNEG) (Mathworks, Natick, MA) that minimizes the sum of the squared deviations (χ^2) between observed and calculated values of $I(t)$, subject to the non-negativity constraint. This algorithm produces the narrowest $P(k)$ distributions and smallest values of χ^2 with relatively few nonzero components. The second method is based

on the Maximum Entropy (ME) theory. The information theory proposes that the least biased solution to the inversion problem is to minimize χ^2 and maximize the breadth of $P(k)$ (48). This regularization condition can be met by maximizing the Shannon-Jaynes entropy of the rate-constant distribution $\left\{S = -\sum_k P(k) \ln[P(k)]\right\}$ while satisfying the non-negativity constraint. ME fitting generated stable and reproducible numerical inversions of the kinetics data. The balance between minimization and entropy maximization is evaluated by the L-curve analysis, which yielded upper limits for the widths of $P(k)$ consistent with experimental data. The $P(k)$ distributions from ME fitting were broader than those obtained with LSQ fitting, but exhibited maxima at similar locations.

Both methods were used to generate the decay rate distribution $P(k)$. A coordinate transformation using the Förster relation (Eq. 1) was then used to convert the probability distribution of the decay rates k to the donor-acceptor distances r , thus generating the donor-acceptor distance distribution $P(r)$.

$$r = R_0 \left(\frac{k}{k_0} - 1 \right)^{1/6} \quad (1)$$

The Förster radius, R_0 , for the DACM/BODIPY-FL pair is $\sim 47 \text{ \AA}$. The value of k_0 was obtained from donor-only measurements, which gave a nearly single-exponential ($>90\%$) fluorescence decay kinetics for all three positions in this study. At distances larger than $1.5 R_0$, energy transfer does not take place efficiently, whereas at distances $< \sim 13 \text{ \AA}$, the Förster model does not reliably describe FRET kinetics. Therefore, our TR-FRET measurements can provide information about donor-acceptor distances only in the range of $13 - 70 \text{ \AA}$.

Fluorescence anisotropy measurements. Anisotropy measurements used excitation and emission wavelengths of 380 nm and 470 nm for DACM and 450 nm and 518 nm for BODIPY, respectively.

Fluorescence anisotropy was calculated according to Eq. 2:

$$R = \frac{(I_{VV} - G \times I_{VH})}{(I_{VV} + 2G \times I_{VH})}, \quad (2)$$

in which I_{VV} and I_{VH} are the vertically and horizontally polarized emission intensities when the sample is vertically excited; G is the grating factor that corrects for the wavelength response to polarization of the emission optics and detectors, defined as $G = I_{HV}/I_{HH}$, where I_{HV} and I_{HH} are the vertically and horizontally polarized emission intensities when the sample is horizontally excited.

Contribution of dipole orientation and fluorophore linkers to distance distribution.

Fluorescence anisotropy measurements showed that both the donor and acceptor fluorophores exhibited low anisotropy values comparable to the free dye when they were incorporated into the proteins. This strongly suggests that the labeled fluorophores are relatively free rotamers with randomized orientations. Hence, the orientation factor, κ^2 , can be approximated by $\langle \kappa^2 \rangle = 2/3$. In addition, the distance distribution can be widened and/or shifted by the fluorophore linkers. For DACM, the linker length is short and very rigid, and thus the primary contribution of the linker is to shift the measured distances by $\sim 5 \text{ \AA}$. On the other hand, BODIPY-FL has a long (6 carbon bonds) and flexible linker that will widen the distance distribution. This effect was estimated as one effective Gaussian chain with the parameter, $r^{\text{linker}} = \sqrt{L \times l_p}$, in which L and l_p are the contour and persistence lengths of the fluorescence linker, respectively (49). This yielded an estimated r^{linker} value of $\sim 7 \text{ \AA}$ for BODIPY-FL.

GTPase assay. The assay to measure the stimulated GTP hydrolysis reaction between SRP and SR was performed and analyzed as described (42). Briefly, reactions were carried out in SRP buffer in the presence of a small, fixed amount of SRP (100–200 nM), varying amounts of SR, and saturating GTP (100 – 200 μM). The observed rate constants (k_{obsd}) were plotted against SR concentration and fit to Eq. 3,

$$k_{obsd} = k_{cat} \times \frac{[SR]}{K_m + [SR]}, \quad (3)$$

in which k_{cat} is the maximal rate constant at saturating SR concentrations, and K_m is the concentration required to reach half saturation. Because k_{cat} is at least 100-fold faster than the rate of SRP•SR complex disassembly, the rate constant k_{cat}/K_m in this assay is rate-limited by and therefore equal to the rate of stable SRP-SR complex formation (42). No DTT was present in the reactions involving spin-labeled proteins.

Docking. The ClusPro 2.0 docking server was used to generate docking models for the early intermediate (50). This program was chosen because it emphasizes the number of energy-preferred structures in the docking cluster, and is therefore particularly suitable to generate an ensemble of conformations for the early intermediate. During the docking, *E. coli* Ffh was set as a static receptor, while *E. coli* FtsY was set as a ligand that searched for the best docking position with the receptor. The initial docking positions were generated by the Fast Fourier Transform method without using the FRET distances as constraints; and the resulting docking positions were clustered according to their root mean squares deviations. The best energy conformations were sorted as clusters via a filter that was set to an energy function that favors electrostatic interactions. The ranking of the clusters was determined by the number of structures that each cluster contained. The top five clusters had 89, 88, 65, 59, and 46 structures, respectively. The top two clusters, named ‘G’ and ‘N’, were chosen for further analyses.

Brownian dynamics. BrownDye was used for Brownian Dynamics calculations (30). APBS was used to calculate the electrostatic potentials (28). Partial atomic charges and atomic radii were assigned from the PARSE parameter set. The dielectric constants were assigned to be 4 in the protein interior and 78 in the exterior. Grids were assigned with dimensions of 193 × 193 × 193 points. Temperature was set to 298 K and ionic strength was set to 100 mM. Brownian dynamics trajectories were started at a minimum intermolecular separation that still gave spherically symmetric forces. The number of trajectories to estimate the association rate constants varied from

40,000 to 100,000 depending on how fast the rates were. The reaction criterion was specified by the atom-contact pairs defined by the structure of the complex. All the intermolecular nitrogen-oxygen pairs within 0.55 nm were considered as within the reaction criterion. A series of simulations with different levels of reaction criteria was generated by systematically tuning the required atom-contact number from 3 to 7. Three structures were used for this analysis to obtain the association rate constants: the central structure of the ‘G’ cluster, the central structure of the ‘N’ cluster, and the crystal structure of the stable complex.

EPR. EPR spectra were acquired with a 9.4 GHz (X-band) Bruker EMX EPR spectrometer with an ER 4119HS cavity at 20-23 °C. 40% glycerol was present in all samples to eliminate the global tumbling motion of proteins. All scans were carried out using a microwave power of 5 mW, a modulation amplitude of 2 gauss and a magnetic field sweep width of 100 gauss. The central linewidth of EPR spectra was the same at microwave powers of 0.2 – 5 mW. Averaged spectra were obtained from 32 – 64 scans and background signals were subtracted.

EPR conditions. EPR measurements were carried out in SRP buffer [50 mM KHEPES, pH 7.5, 150 mM KOAc, 2 mM Mg(OAc)₂, 2 mM DTT, 0.01% Nikkol] to determine the local mobility of twenty three spin-labeled FtsY mutants in apo-FtsY, in the early intermediate, and in the stable complex. For apo-FtsY, 75-100 μM spin-labeled protein was used to obtain the EPR spectra. The early intermediate was formed by mixing 30 μM spin-labeled FtsY with 90 μM SRP in the presence of GDP. Based on the affinity of the early intermediate ($K_d \sim 4\text{-}10 \mu\text{M}$) (21), >90% of labeled FtsY formed the early complex with SRP under these conditions. The stable complex was formed by mixing 30 μM spin-labeled-FtsY with 60 μM SRP in the presence of GMPPNP. Over 99% of labeled FtsY formed a stable complex with SRP under these conditions, according to the K_d values of the stable complex of ~16–30 nM (21).

Steady-state fluorescence. All measurements were carried out at 25 °C in SRP buffer [50 mM KHEPES, pH 7.5, 150 mM KOAc, 2 mM Mg(OAc)₂, 2 mM DTT, 0.01% Nikkol] on a Fluorolog-

3-22 spectrofluorometer (Jobin Yvon, Edison, NJ), using an excitation wavelength of 380 nm and an emission wavelength of 470 nm. FRET efficiency was calculated as described (21). To compare the relative equilibrium stabilities of the early intermediates formed by different SR mutants, 4 μ M BODIPY-FL labeled SR was incubated with 1 μ M DACM-labeled SRP in the absence of GTP. As formation of the early intermediate is rapid but has a high K_d (4 – 10 μ M), the FRET value at the sub-saturating SR concentration provided a sensitive measure of the changes in its stability. For representative mutants, equilibrium titrations were carried out. The data were fit to Eq. 4

$$E = E_1 \times \frac{[\text{SR}]}{K_d + [\text{SR}]}, \quad (4)$$

in which E_1 is the FRET end point with saturating SR, and K_d is the equilibrium dissociation constant of the early intermediate.

References

1. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291-294.
2. Janin J (1997) The kinetics of protein-protein recognition. *Proteins* 28:153-161.
3. Schreiber G (2002) Kinetic studies of protein-protein interactions. *Curr Opin Struct Biol* 12:41-47.
4. Schreiber G, Haran G, Zhou HX (2009) Fundamental Aspects of Protein-Protein Association Kinetics. *Chem Rev*.
5. Ubbink M (2009) The courtship of proteins: understanding the encounter complex. *FEBS Lett* 583:1060-1066.
6. Koshland DE (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A* 44:98-104.
7. Ma B, Kumar S, Tsai CJ, Nussinov R (1999) Folding funnels and binding mechanisms. *Protein Eng* 12:713-720.

8. Schreiber G, Shaul Y, Gottschalk KE (2006) Electrostatic design of protein-protein association rates. *Methods Mol Biol* 340:235-249.
9. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* 5:789-796.
10. Shoemaker BA, Portman JJ, Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci U S A* 97:8868-8873.
11. Papoian GA, Ulander J, Wolynes PG (2003) Role of water mediated interactions in protein-protein recognition landscapes. *J Am Chem Soc* 125:9170-9178.
12. Papoian GA, Wolynes PG (2003) The physics and bioinformatics of binding and folding- an energy landscape perspective. *Biopolymers* 68:333-349.
13. Tang C, Iwahara J, Clore GM (2006) Visualization of transient encounter complexes in protein-protein association. *Nature* 444:383-386.
14. Volkov AN, Worrall JA, Holtzmann E, Ubbink M (2006) Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR. *Proc Natl Acad Sci U S A* 103:18945-18950.
15. Fawzi NL, Doucleff M, Suh JY, Clore GM (2010) Mechanistic details of a protein-protein association pathway revealed by paramagnetic relaxation enhancement titration measurements. *Proc Natl Acad Sci U S A* 107:1379-1384.
16. Crowley PB, Ubbink M (2003) Close encounters of the transient kind: protein interactions in the photosynthetic redox chain investigated by NMR spectroscopy. *Acc Chem Res* 36:723-730.
17. Xu X, *et al.* (2008) Dynamics in a pure encounter complex of two proteins studied by solution scattering and paramagnetic NMR spectroscopy. *J Am Chem Soc* 130:6395-6403.

18. Zhang X, Schaffitzel C, Ban N, Shan SO (2009) Multiple conformational switches in a GTPase complex control co-translational protein targeting. *Proc Natl Acad Sci U S A* 106:1754-1759.
19. Zhang X, Rashid R, Wang K, Shan SO (2010) Sequential checkpoints govern substrate selection during cotranslational protein targeting. *Science* 328:757-760.
20. Egea PF, *et al.* (2004) Substrate twinning activates the signal recognition particle and its receptor. *Nature* 427:215-221.
21. Zhang X, Kung S, Shan SO (2008) Demonstration of a multistep mechanism for assembly of the SRP x SRP receptor complex: implications for the catalytic role of SRP RNA. *Journal of Molecular Biology* 381:581-593.
22. Shan SO, Stroud RM, Walter P (2004) Mechanism of association and reciprocal activation of two GTPases. *Plos Biology* 2:1572-1581.
23. Crane JM, Lilly AA, Randall LL (2010) Characterization of interactions between proteins using site-directed spin labeling and electron paramagnetic resonance spectroscopy. *Methods Mol Biol* 619:173-190.
24. McHaourab HS, Lietzow MA, Hideg K, Hubbell WL (1996) Motion of spin-labeled side chains in T4 lysozyme. Correlation with protein structure and dynamics. *Biochemistry* 35:7692-7704.
25. Zhang X, Lee SW, Zhao L, Xia T, Qin PZ (2010) Conformational distributions at the N-peptide/boxB RNA interface studied using site-directed spin labeling. *RNA* 16:2474-2483.
26. Kimura T, Lee JC, Gray HB, Winkler JR (2009) Folding energy landscape of cytochrome cb562. *Proc Natl Acad Sci U S A* 106:7834-7839.
27. Focia PJ, Gawronski-Salerno J, Coon JS, Freymann DM (2006) Structure of a GDP : AIF4 complex of the SRP GTPases Ffh and FtsY, and identification of peripheral nucleotide interaction site. *Journal of Molecular Biology* 360:631-643.

28. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98:10037-10041.
29. Comeau SR, *et al.* (2007) ClusPro: performance in CAPRI rounds 6-11 and the new server. *Proteins* 69:781-785.
30. Ermak DL, Mccammon JA (1978) Brownian Dynamics with Hydrodynamic Interactions. *Journal of Chemical Physics* 69:1352-1360.
31. Estrozi LF, Boehringer D, Shan SO, Ban N, Schaffitzel C (2011) Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. *Nat Struct Mol Biol* 18:88-90.
32. Shen K, Shan SO (2010) Transient tether between the SRP RNA and SRP receptor ensures efficient cargo delivery during cotranslational protein targeting. *Proc Natl Acad Sci U S A* 107:7698-7703.
33. Levy Y, Onuchic JN, Wolynes PG (2007) Fly-casting in protein-DNA binding: frustration between protein folding and electrostatics facilitates target recognition. *J Am Chem Soc* 129:738-739.
34. Miyashita O, Onuchic JN, Okamura MY (2004) Transition state and encounter complex for fast association of cytochrome c2 with bacterial reaction center. *Proc Natl Acad Sci U S A* 101:16174-16179.
35. Jones S, Thornton JM (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 93:13-20.
36. Schreiber G, Fersht AR (1996) Rapid, electrostatically assisted association of proteins. *Nat Struct Biol* 3:427-431.
37. Vijayakumar M, *et al.* (1998) Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. *J Mol Biol* 278:1015-1024.

38. Kiel C, Selzer T, Shaul Y, Schreiber G, Herrmann C (2004) Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proc Natl Acad Sci U S A* 101:9223-9228.
39. Harel M, Spaar A, Schreiber G (2009) Fruitful and futile encounters along the association reaction between proteins. *Biophys J* 96:4237-4248.
40. Pande VS, Rokhsar DS (1998) Is the molten globule a third phase of proteins? *Proc Natl Acad Sci U S A* 95:1490-1494.
41. Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. *Science* 267:1619-1620.
42. Peluso P, Shan SO, Nock S, Herschlag D, Walter P (2001) Role of SRP RNA in the GTPase cycles of Ffh and FtsY. *Biochemistry* 40:15224-15233.
43. Schaffitzel C, Ban N (2007) Generation of ribosome nascent chain complexes for structural and functional studies. *Journal of Structural Biology* 158:463-471.
44. Bornemann T, Jockel J, Rodnina MV, Wintermeyer W (2008) Signal sequence-independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. *Nat Struct Mol Biol* 15:494-499.
45. Lee JC, Engman KC, Tezcan FA, Gray HB, Winkler JR (2002) Structural features of cytochrome c' folding intermediates revealed by fluorescence energy-transfer kinetics. *Proc Natl Acad Sci U S A* 99:14778-14782.
46. Beals JM, Haas E, Krausz S, Scheraga HA (1991) Conformational studies of a peptide corresponding to a region of the C-terminus of ribonuclease A: implications as a potential chain-folding initiation site. *Biochemistry* 30:7680-7692.
47. Beechem JM, Haas E (1989) Simultaneous determination of intramolecular distance distributions and conformational dynamics by global analysis of energy transfer measurements. *Biophys J* 55:1225-1236.

48. Istratov AA, Vyvenko OF (1999) Exponential analysis in physical phenomena. *Review of Scientific Instruments* 70:1233-1257.
49. Laurence TA, Kong X, Jager M, Weiss S (2005) Probing structural heterogeneities and fluctuations of nucleic acids and denatured proteins. *Proc Natl Acad Sci U S A* 102:17348-17353.
50. Kozakov D, *et al.* (2010) Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins* 78:3124-3130.

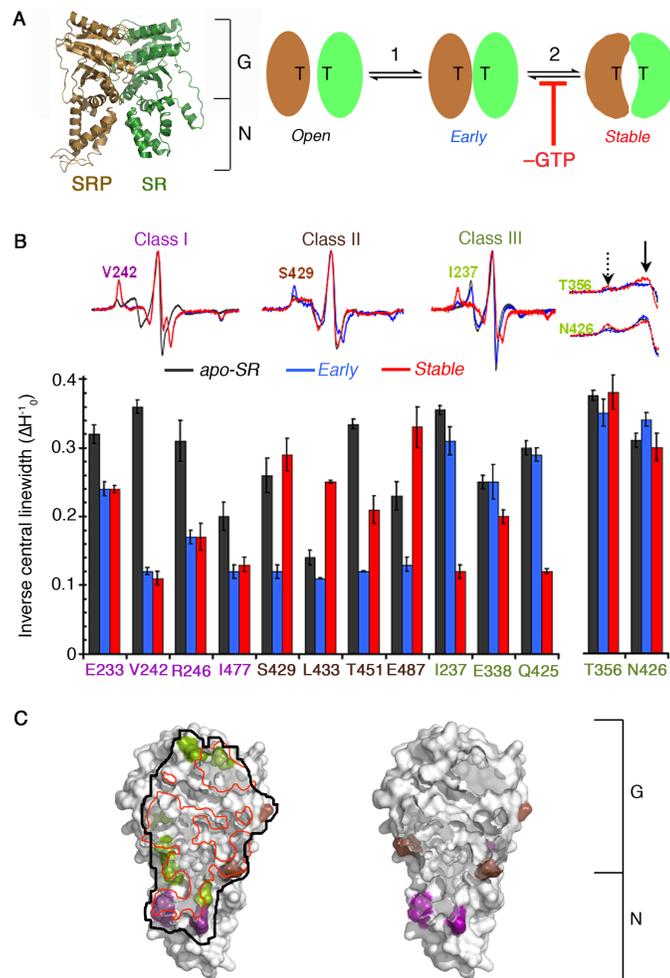


Fig. A-1 Mapping the interaction interface of the SRP•SR complexes using EPR spectroscopy. (A) Left: crystal structure of the SRP•SR NG-domain complex (1JR9). Right: a multi-step mechanism for SRP-SR complex assembly involving formation of an early intermediate (step 1), and rearrangement to form the stable complex (step 2). Removal of GTP stalls the complex at the early intermediate stage. T denotes GTP. (B) Nitroxide spin probes labeled at specific SR residues change mobility upon formation of the early intermediate (blue), the stable complex (red), or both. The different classes of spin probe mobility changes are defined in the text. Black denotes the apo-formed SR. (C) Interaction surface of the stable complex (left) and early intermediate (right) mapped by EPR. ‘N’ and ‘G’ denote the N- and G-domains of SRP and SR, respectively. The red line outlines the interaction surface in the co-crystal structure (20) of the stable complex.

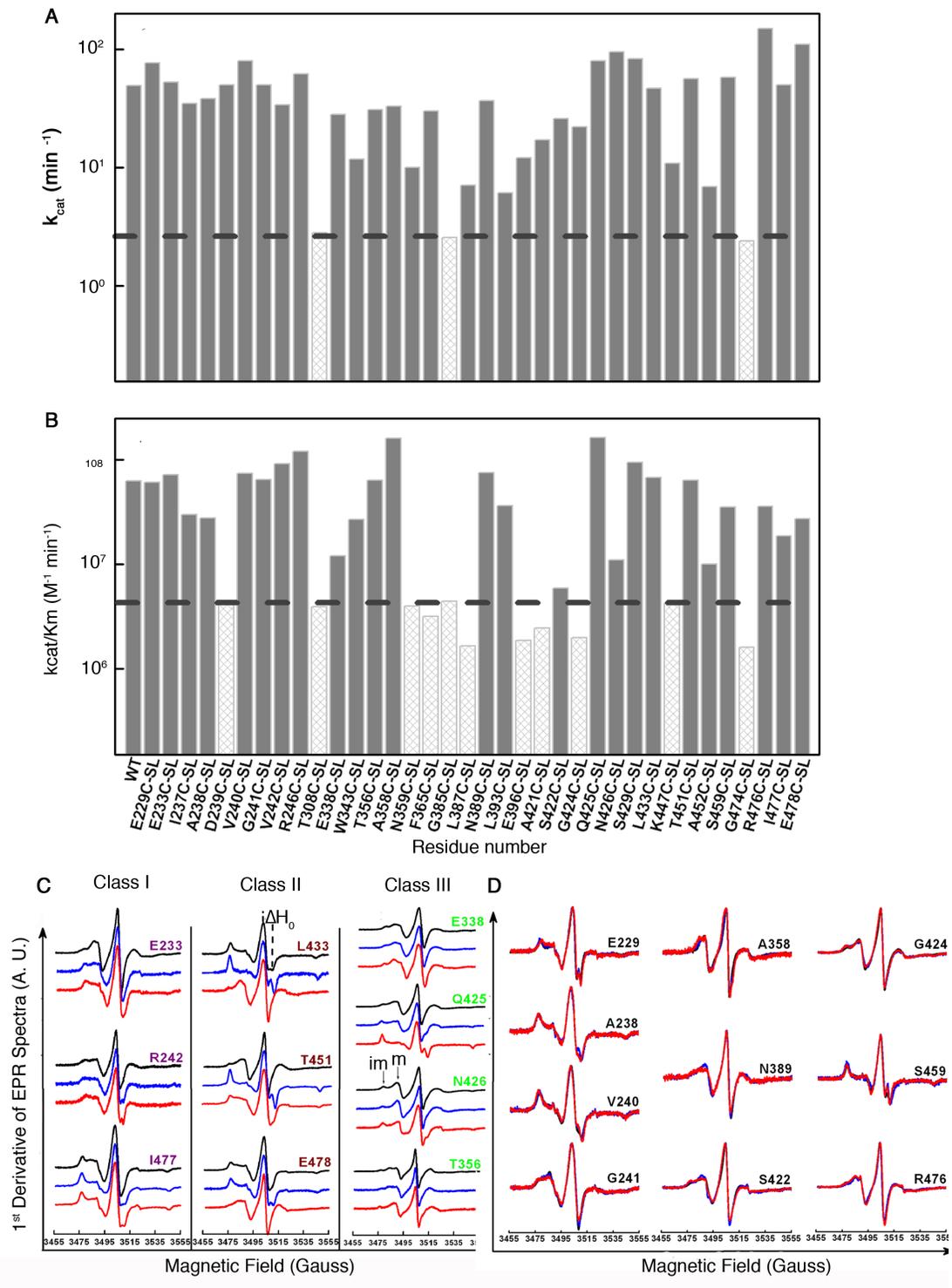


Fig. A-2 The mobility of spin labels on SR changed upon formation of the early intermediate, stable complex, or both. (A and B) Spin-labeled SR were screened using the GTPase assay. The activities of spin-labeled (SL) SR's in interaction with SRP were analyzed using the GTPase assay (see

Supplementary Methods). Two kinetic parameters were assessed: the GTPase rate constants of the SRP-SR complex (k_{cat} in *A*) and the association rate constants for stable SRP-SR complex assembly (as determined by k_{cat}/K_m in *B*; see Supplementary Methods). Spin-labeled SR's that were defective in either property by a factor of 5 or more were not used for EPR studies (open bars). Spin-labeled SR's that were functional in interacting with SRP (grey bars) were used for EPR measurements for either the early intermediate or the stable complex. (*C*) Spectra of additional spin probes in SR changed mobility upon formation of either the early intermediate or the stable complex. Three different classes were defined in the text based on probe mobility changes. The mobility of spin label was analyzed from the central line width (ΔH_0) and the breadth of the spectra, and summarized in Figure 1*B*. Black, blue, and red denote the free protein, the early intermediate, and the stable complex, respectively. (*D*) EPR spectra of spin probes in SR that exhibited no significant changes in mobility upon formation of either complex. Color-coding is the same as in *A*.

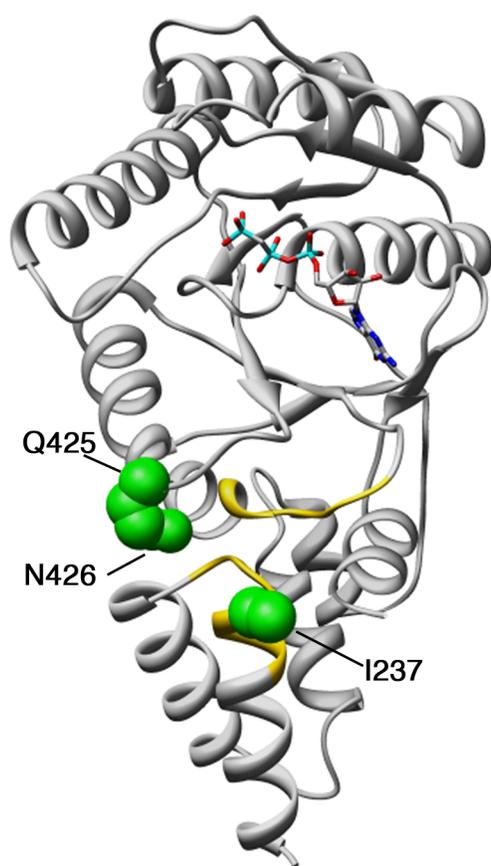


Fig. A-3 Residues I237, Q425, and N426 (green), which changed EPR spectra specifically in the stable complex, are at the conserved motifs (yellow) that mediate N-G domain rearrangement.

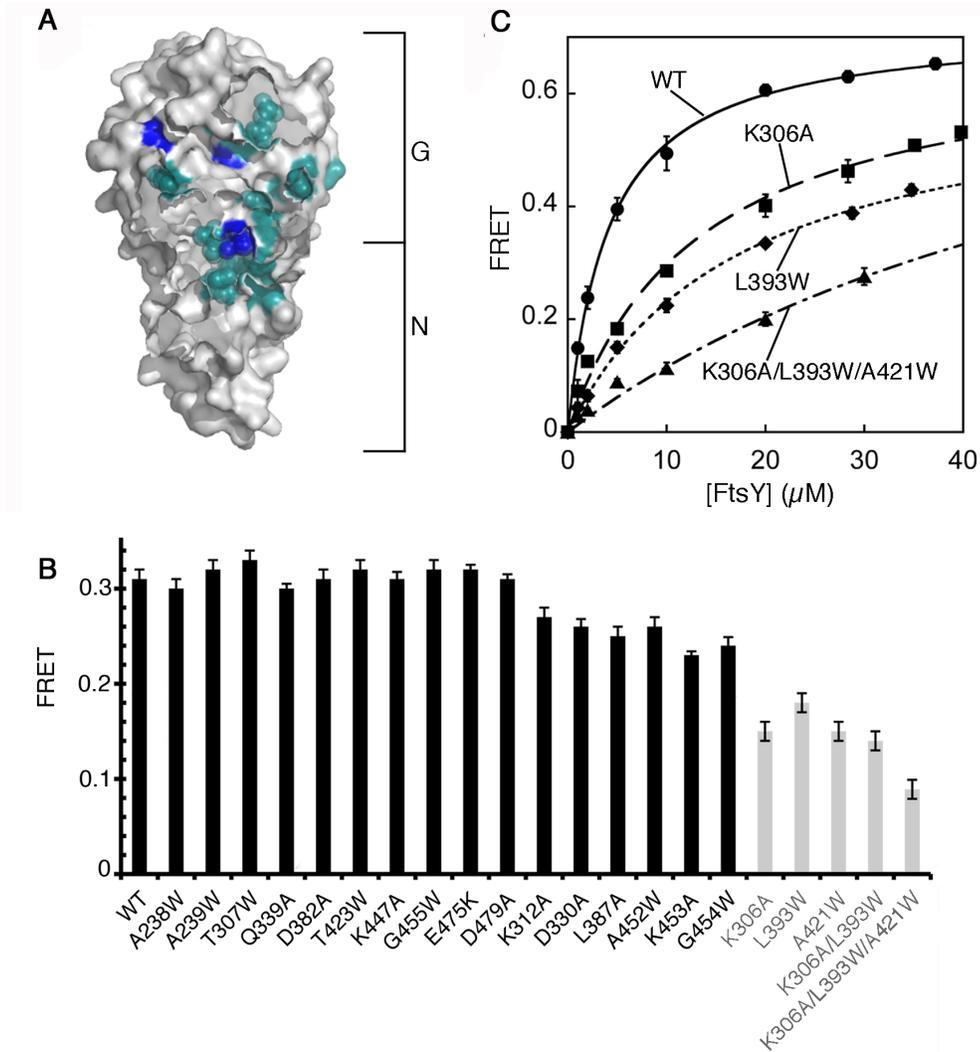


Fig. A-4 Mutations that disrupt the stable complex did not significantly affect the early intermediate. (A) Positions of SR mutations (cyan and blue) studied herein are shown in the surface representation of the SR. The three moderately defective mutants are highlighted in blue. (B) The stability of the early intermediate is insensitive to many mutations that disrupt the stable complex. (C) The stabilities of the early intermediates formed by mutant SRs were determined by equilibrium titrations. Nonlinear fits gave K_d values of 4.1 μM for wild-type SR, 13.2 μM for SR (K306A), 17.3 μM for SR (L393W), and 31.3 μM for SR (K306A:L393W:A421W).

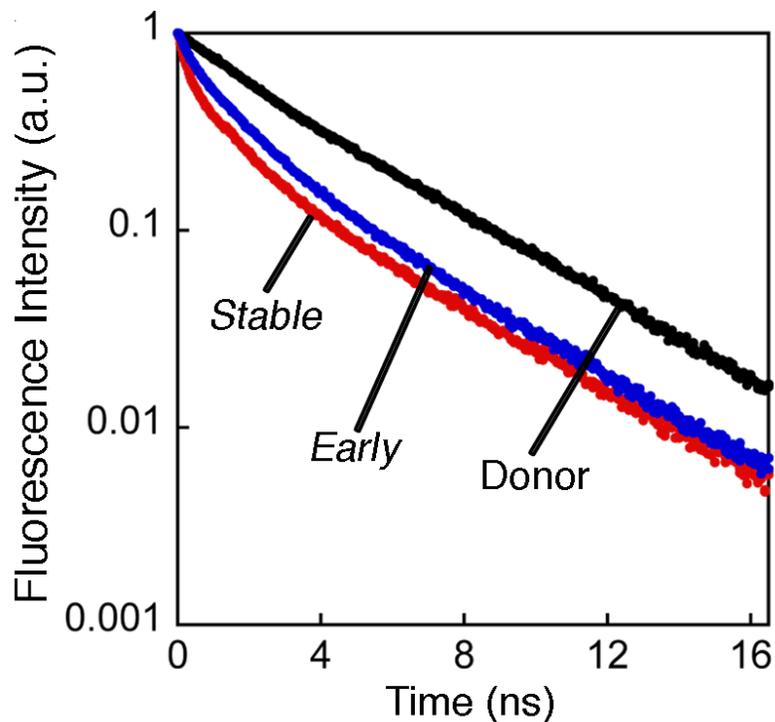


Fig. A-5 Fluorescence decay of donor (DACM)-labeled at SRP (C76) under different experimental conditions. The black, blue, and red curves represent the decay curves for donor-only, the early intermediate, and the stable complex, respectively. The linear decay of the donor-only sample could be described by a single decay rate constant. In contrast, the decay curves in both the early intermediate and stable complex deviated from linearity and were described by multiple decay rate constants.

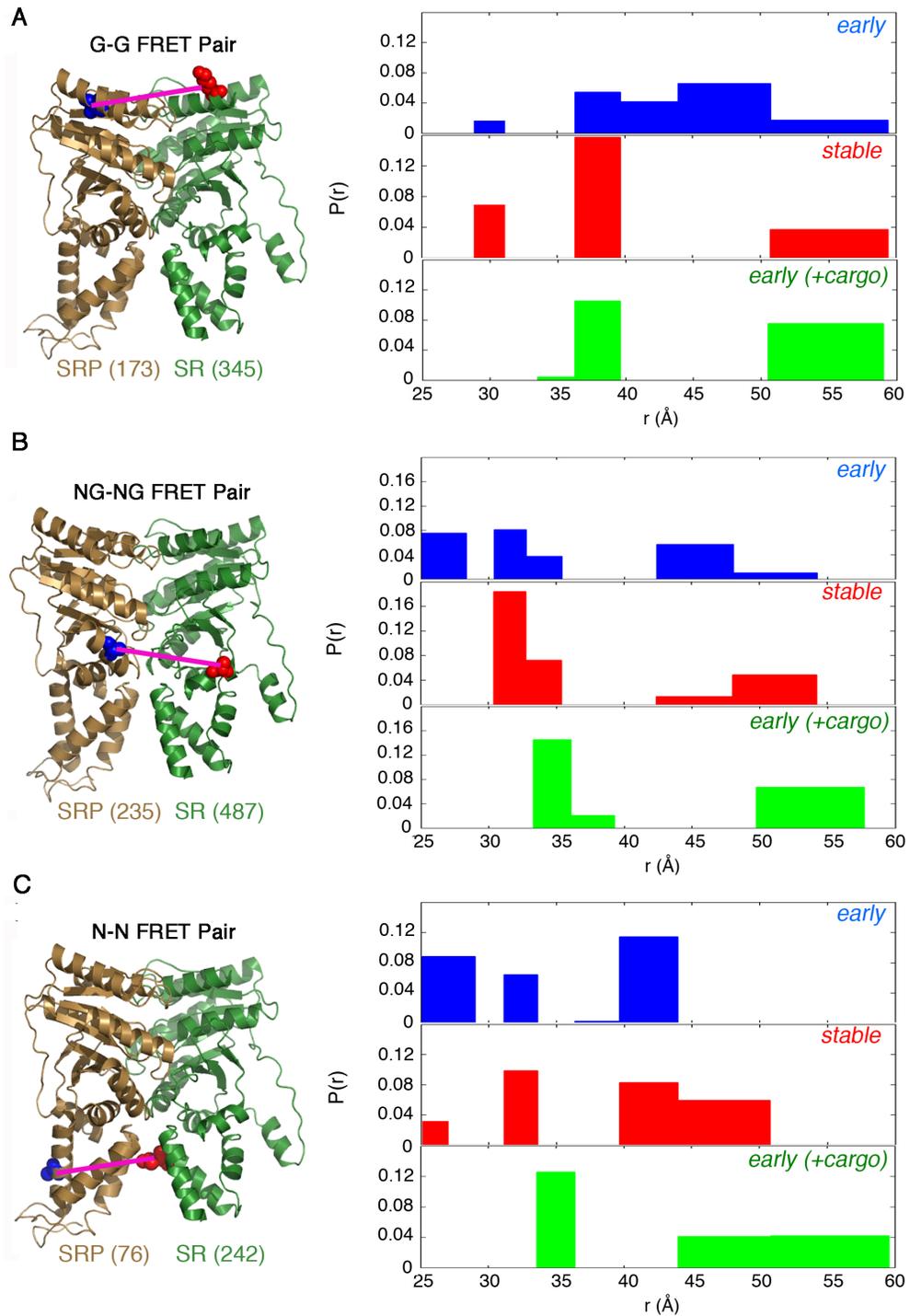


Fig. A-6 Distance distributions derived from least-squares analyses of the TR-FRET data for each FRET pair in the early intermediate (blue), stable complex (red), and early intermediate bound with cargo (green).

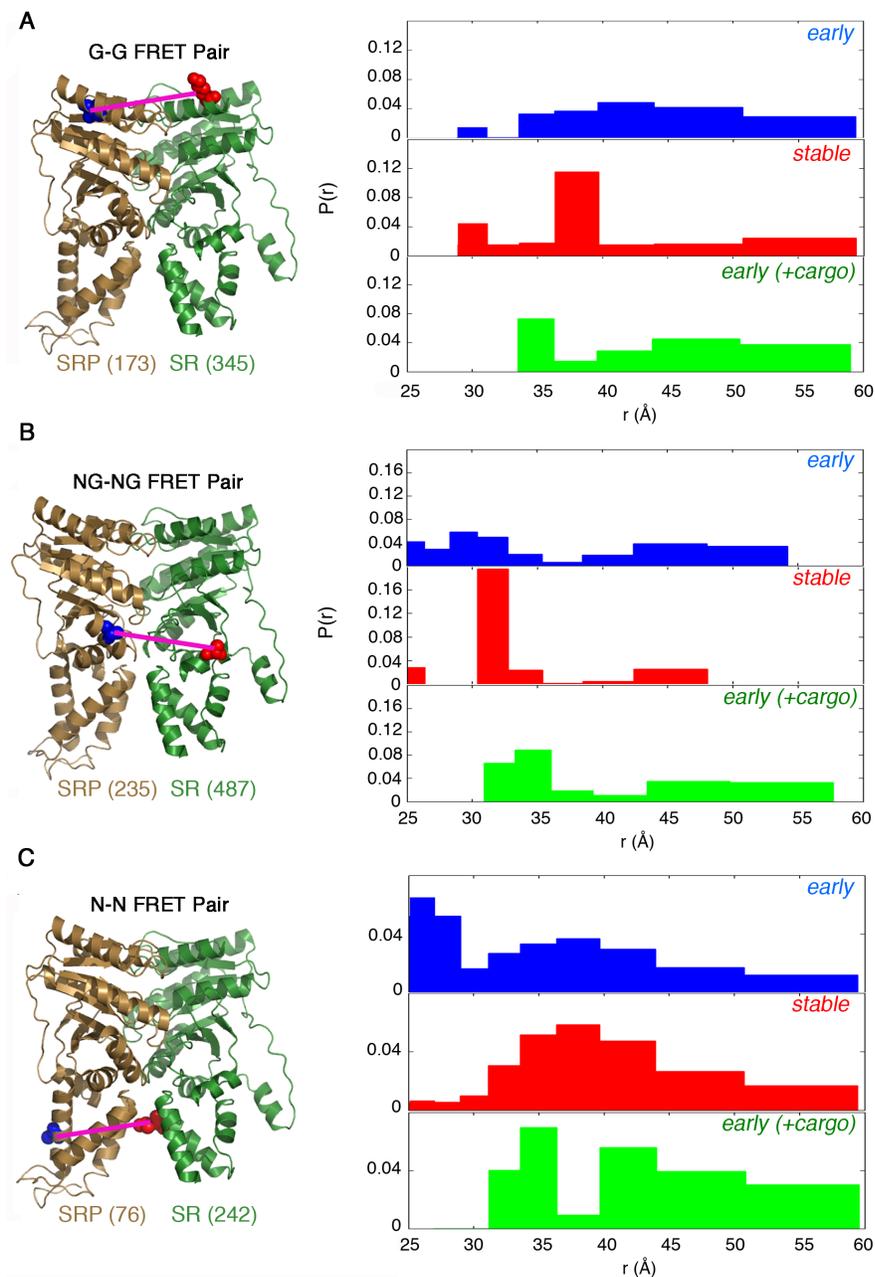


Fig. A-7 Conformational distribution of the early intermediate is broad, and is restricted by formation of the stable complex or the cargo. Left, positions of the G-G (A), NG-NG (B), and N-N (C) FRET pairs in the stable SRP•SR NG-domain complex. Right, FRET distance distributions, $P(r)$, for each FRET pair in the early intermediate (blue), stable complex (red), and the early intermediate bound to cargo (green), as derived from maximal entropy analyses of the TR-FRET data.

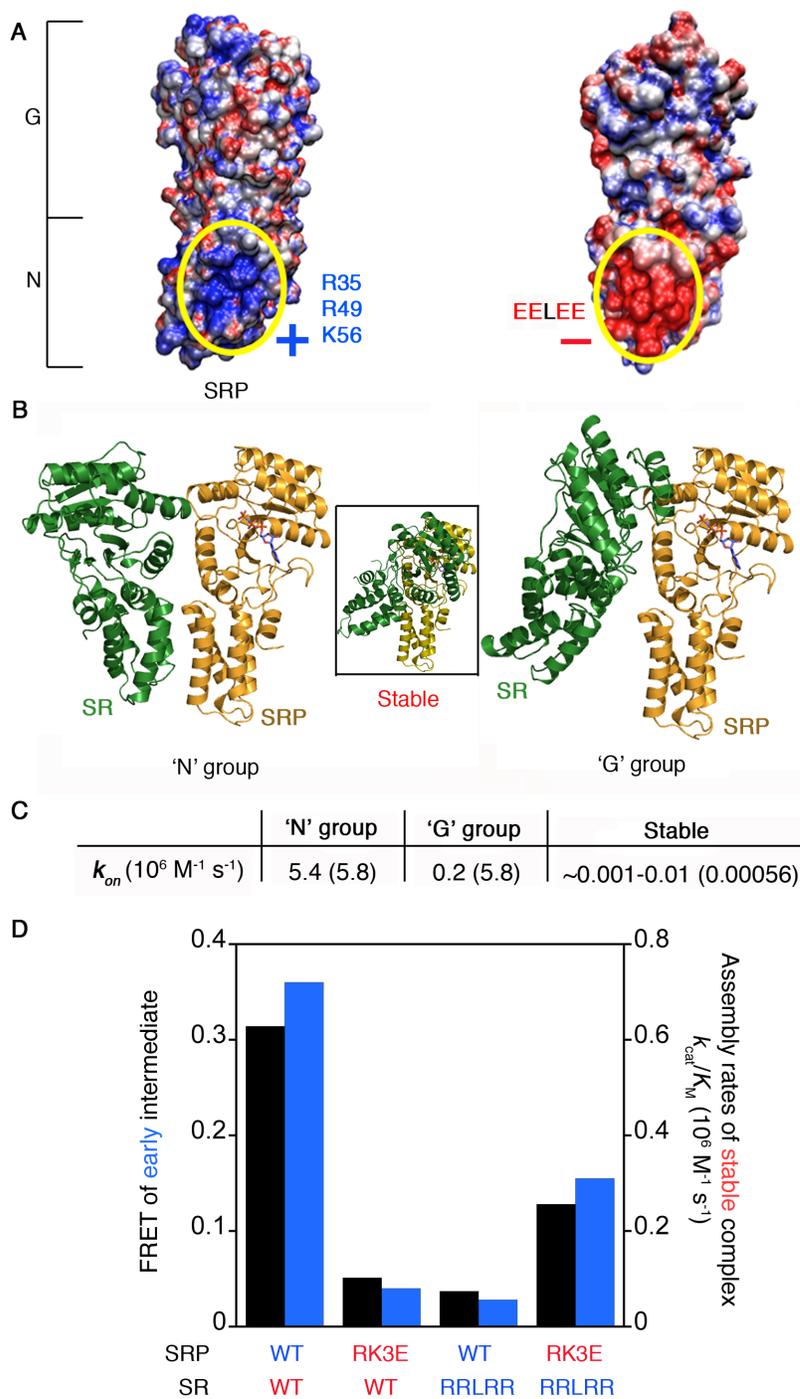


Fig. A-8 Electrostatic interactions between the N-domains of SRP and SR stabilize the early intermediate and accelerate stable complex assembly. (A) The SRP and SR N-domains contain complementarily charged surfaces. (B) Molecular docking simulation generated two groups of conformations ('N' and 'G') for the early intermediate. For comparison, the inset in the middle

shows the structure of the stable complex with the SRP NG-domain aligned in the same orientation. The nucleotide bound to SRP is shown in CPK coloring. (C) The association rate constants predicted from Brownian Dynamics calculations for formation of the early intermediate in the 'N' or 'G' group, and for the stable complex. The experimentally measured rate constants are in parentheses. (D) Charge complementarity between the N-domains is critical to the stability of the early intermediate (black bars) and the kinetics of stable complex assembly (blue bars), determined using the FRET and GTPase assays, respectively, for the wildtype proteins (WT:WT), wildtype SRP and mutant SR (WT:RRLRR), mutant SRP and wildtype SR (RK3E:WT), and the charge reversal SRP and SR mutant pair (RK3E:RRLRR). The kinetic constants were derived from the data in Figure S5D. FRET efficiency in the early intermediate was recorded at 5 μ M FtsY, at which concentration the FRET value is most sensitive to changes in the stability of the early intermediate.

N-domain (left), and the RRLRR mutation in SR generated a moderately positive electrostatic potential in the SR N-domain (right). (C) The stabilities of the early intermediates formed by mutant SRP and SR's were determined by equilibrium titrations. Nonlinear fits of data to eq. 1 (main text) gave K_d values of 4.0 μM for WT:WT (wild-type SRP and SR), 50.1 μM for R2E:WT [mutant SRP (R2E) and wild-type SR], and 20.1 μM for R2E:RRLRR [mutants SRP (R2E) and SR (RRLRR)]. (D) The kinetics of stable complex assembly. Nonlinear fits of the data gave k_{cat}/K_m values of 0.72×10^6 , 0.056×10^6 , 0.080×10^6 , and $0.31 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$, respectively, for the interaction between the wildtype proteins (WT:WT), wildtype SRP and mutant SR (WT:RRLRR), mutant SRP and wildtype SR (RK3E:WT), and the charge reversal SRP and SR mutants (RK3E:RRLRR).

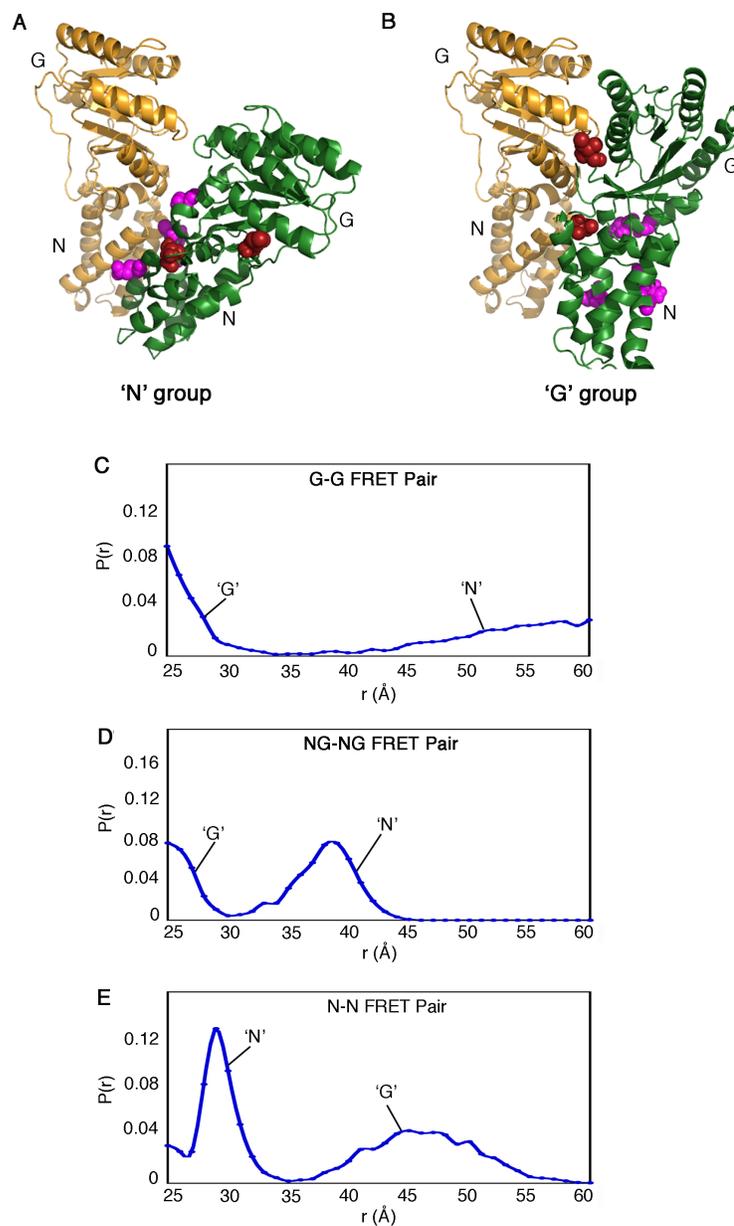


Fig. A-10 The 'N' and 'G' groups represent possible conformations within the ensemble of the early intermediate (A-B) Spin probes that changed mobility upon formation of the early intermediate are close to the interaction surface of either the 'N' (magenta residues) or the 'G' (red residues) group. The SRP NG-domain is in gold, and the SR NG-domain is in green. (C-E) Distance distributions of the three pairs of FRET probes predicted by a combination of the docking structures in the 'N' and 'G' groups.

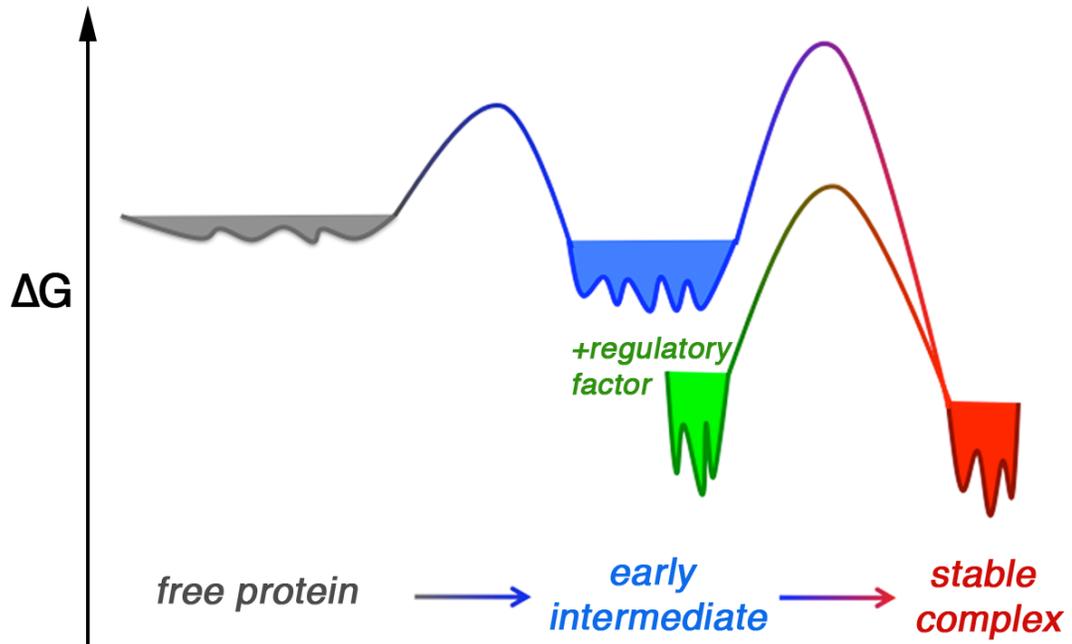


Fig. A-11 Model of free energy landscapes for the protein assembly process. The conformational space is broad for the free proteins (grey) and the early intermediate (blue), but becomes more restricted in the stereospecific stable complex (red) or when SRP is bound to the cargo (green).