Engineering Gene Targeting Reagents Through Computational Design and Directed Evolution of Protein-DNA Interactions

Summer B. Thyme

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2012

Reading Committee:
David Baker, Chair
Philip Bradley
Barry Stoddard

Program Authorized to Offer Degree:
Department of Biochemistry

University of Washington

Abstract

Engineering Gene Targeting Reagents Through Computational Design and Directed Evolution of Protein-DNA Interactions

Summer B. Thyme

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Homing endonucleases have great potential as tools for targeted gene therapy and gene correction. These DNA cleavage enzymes are capable of inducing gene repair or disruption by stimulating DNA repair pathways at a specific location in a genome. However, identifying variants of these enzymes capable of cleaving specific DNA targets of interest is necessary before the widespread use of such technologies is possible. Enzyme engineering should be informed by detailed analyses of the wild-type system. In my early work redesigning the target site specificity of the homing endonuclease I-AniI, it was discovered that, despite the approximate two-fold symmetry of the enzyme-DNA complex, there is almost complete segregation of the interactions responsible for substrate binding and transition state stabilization. This separation of the roles of these domains was revealed by doing kinetic studies on target sites differing by a single base-pair from the wild-type substrate. Computationally redesigned variants of I-AniI that achieved new specificities on one side did so by modulating binding, while redesigns with altered specificities on the other side modulated catalysis. While some computational designs showed successfully altered specificity, many designs targeting other single base-pair substitutions were unsuccessful. The next step of my work involved developing better computational methods in order to improve recapitulation of the experimental data. I made improvements the

ROSETTA program, giving an energy bonus to native-like interactions collected from the RCSB protein databank and developing protocols for sampling diverse design sequences. Despite advancements in both directed evolution and computational methods, protein engineering is challenging and exploring alternative methods for altering endonuclease specificity is necessary. New endonuclease ORFs can be identified due to substantial increases in available sequence data. I have additionally shown that enzyme hybrids can be built using homologue information, by transferring amino acids from homologues onto the I-AniI scaffold, improving activity and generating new specificities. This wealth of information from millions of years of endonuclease evolution will be used to guide and improve current rational engineering methods.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Role of Genome-Specific DNA Cleavage in Biotechnology and Medicine

The ability to generate genome-specific targeted double-stranded breaks has numerous applications in biotechnology and medicine [1,2,3,4]. The repair of these breaks can occur through either homologous recombination or by an end-joining mechanism (Figure 1) [5]. Homologous recombination relies on the presence of a DNA template with homology to the damaged area to repair the break. This process has the potential to provide site-specific modification of genomes, such as the correction of mutations that cause debilitating genetic diseases [1,2]. The alternative pathway of repair, non-homologous end joining (NHEJ), results in deletions and thus frameshifts and gene knockouts. These deletions can be a detrimental side effect of the break when the potential for reading frame maintenance from homologous recombination is necessary, or they can be a feature of the process if gene disruption is a goal. Genome-specific DNA cleavage enzymes can stimulate these repair processes at targeted locations in a genome. However, it is necessary to develop reagents capable of cleaving any target site of interest with high specificity before the use of these technologies can become widespread. An efficient pathway to this goal is to combine methods for identification of DNA-cleaving enzymes with naturally high specificity with rational engineering methods for altering that specificity.

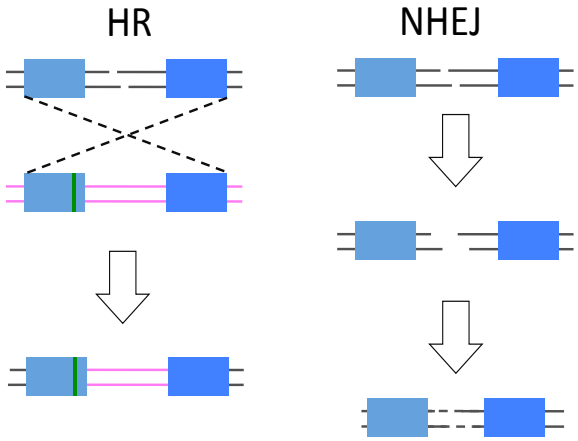
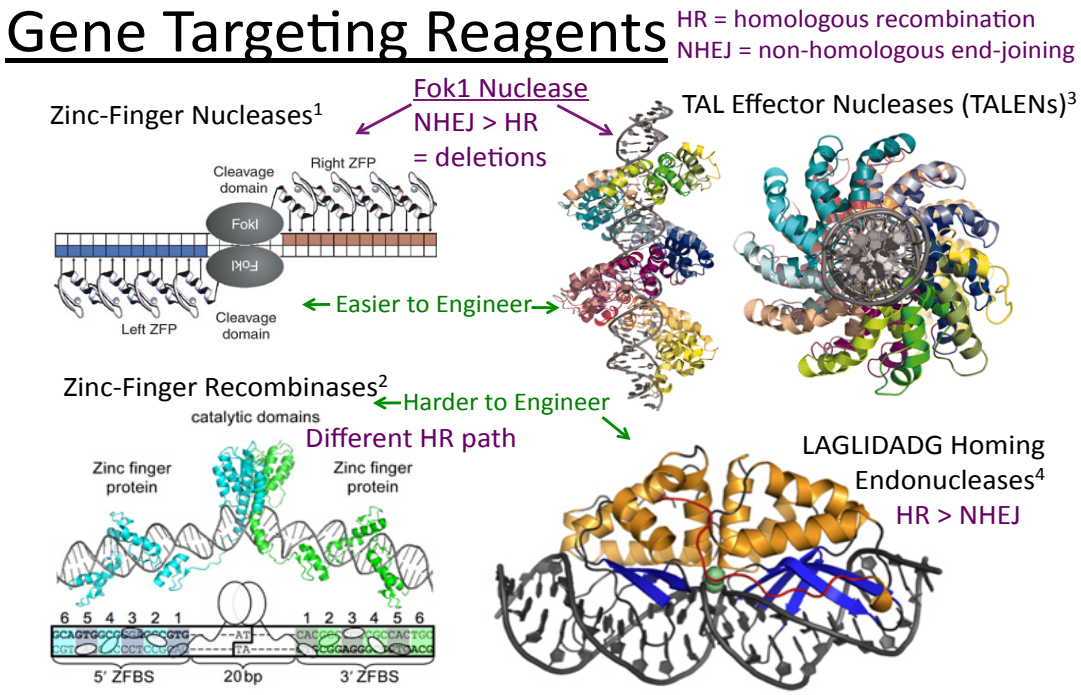


Figure 1. Repair of a double-stranded break. DNA breaks can be repaired through homologous recombination (HR) or non-homologous end-joining (NHEJ). HR utilizes a template to repair the break and incorporates the template information and mutations into the repaired DNA. NHEJ occurs without a template and deletions occur at the break during the repair process.

1.2 The Different Types of Gene Targeting Reagents

Promising platforms for generation of genome-specific cleavage reagents are zinc-finger nucleases (ZFNs) [6,7], TALE nucleases (TALENs) [8,9], zinc-finger recombinases (ZFRs) [10], and LAGLIDADG homing endonucleases or meganucleases [11,12]. These four types of targeting reagents vary in how easy it is to modulate their specificity and how they interact with the DNA damage pathways inside a cell (Figure 2). Both ZFNs and TALENs are very



1. Miller, *et. al.*, Nat. Biotech (2007) 2. Gordley, *et. al.*, (Barbas lab) PNAS (2009) 3. Mak, *et. al.* (Bradley & Stoddard labs) Science (2011) 4. Stoddard lab

Figure 2. Gene targeting reagents. These proteins differ in how easier they are to engineer and how the type of genomic repair that they induce. The reagents that are built from a separable DNA-binding domain and Fok1 nuclease domain are the zinc-finger nucleases [7] and TAL effector nucleases [9]. These two classes of proteins are relatively straightforward to engineer for new specificities, and they tend to induce non-homologous end-joining (NHEJ) repair that results in deletions and gene-knockouts. Zinc-finger recombinases [10] are built from a zinc-finger DNA-binding domain and a recombinase domain, and they are harder to engineer than the Fok1 nuclease derived reagents because the recombinase maintains target DNA specificity that is not well-understood. The last class of enzymes, the LAGLIDADG homing endonucleases [12], are also challenging to engineer because the catalysis is linked to DNA binding and they induce both NHEJ and homologous recombination (HR) outcomes.

engineerable and consist of a DNA-binding domain linked to the same Fok1 nuclease domain. While the technology to generate zinc-finger proteins that bind different DNA targets has been improving over the last decade, the recent identification of the TALE repeat proteins has greatly simplified the problem of building tools with site-specific DNA binding features. The TALEs bind DNA in a straightforward manner, using a code that links two variable interface amino acids residues with a single nucleotide in the target DNA sequence [8,13]. Much of the existing research has been completed with the traditional ZFNs, and pioneering work with these enzymes has shown that specific genome targeting is possible. ZFNs have entered clinical trials for a number of diseases, although the possibility of off-target cleavage has recently dampened progress [14,15,16]. Additionally, one downside to the both nucleases built with a Fok1 cleavage domain is that they tend to induce significant amounts of NHEJ [8]. Reagents for targeting homologous recombination would be highly desirable because genomic sequences can be modified at their native loci without altering their reading frames. Two types of enzymes that show more promise for accomplishing this goal than the Fok1 nuclease fusions are the zinc-finger recombinases (ZFRs) and the homing endonucleases or meganucleases.

ZFRs [10] are newer adaption of the traditional ZFNs [10]. These zinc-finger fusions are not very engineerable because they maintain a recombinase domain attached to the zinc-fingers and that recombinase domain maintains DNA specificity. The specificity of recombinase proteins is not well understood and only recently has there been attempts to modulate it [17]. However, these reagents could be very powerful if the DNA sequence requirements for recombination were known and continued development is worthwhile. A major advantage of these recombinases over the nuclease-driven tools for genome engineering is that recombinases do not depend on the availability of potentially cell-type and cycle specific repair factors for completion of recombination. Additionally, the homologous recombination pathway that is stimulated by nuclease-directed double-stranded breaks – known as single strand annealing or SSA – uses a repair template to copy the template sequence, rather than directly inserting it. Manipulation of DNA with recombinases goes through an alternative pathway where the repair template is physically inserted at the repair locus.

The last family of potential gene targeting reagents that can induce homologous recombination is the LAGLIDADG homing endonucleases or meganucleases. These proteins are significantly harder to engineer than the proteins made of a distinct binding and cleavage

domains because the DNA cleavage event occurs right in the center of the DNA binding region and it's intimately linked to and affected by DNA binding. However, both wild-type and engineered homing endonucleases have also been shown to induce site-specific recombination in cultured mammalian cells, supporting their potential for use as gene therapy reagents [18,19]. The natural role of homing endonucleases is to induce homologous recombination in their hosts (Figure 3). Given that they are evolutionarily optimized for this purpose, it is hypothesized that the ratio of homologous recombination to detrimental end-joining repair pathways will be increased for DSBs generated by homing endonucleases [22]. Specifically, the cleavage of the target site is coupled directly to target site binding and these enzymes are single-turnover [20], suggesting that they are holding onto the ends of the DNA following DNA cleavage, and this feature may minimize the recognition of the breaks by the end-joining repair machinery.

1.3 Properties of the Homing Endonuclease Class of DNA-Cleaving Enzymes

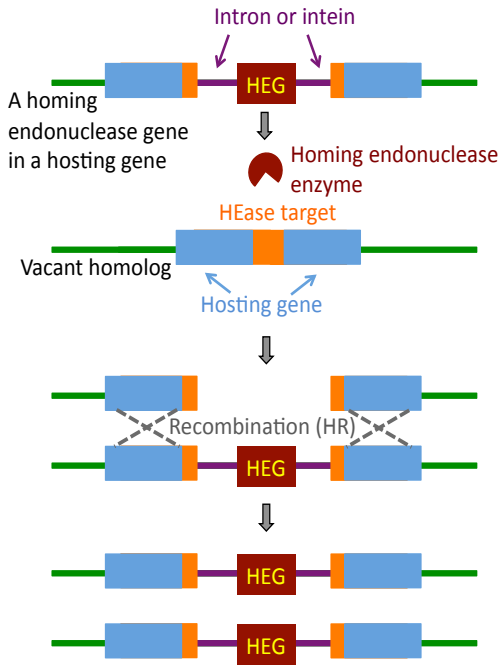


Figure 3. Cycle of homing endonuclease gene propagation. Homing endonuclease genes (HEGs) reside in an intron or intein. The expressed endonuclease enzymes cleave a homologous allele that lacks this intron or intein. This double-stranded break induces transfer of the intron or intein vi homologous recombination and the homing endonuclease gene to the homologous allele.

Homing endonuclease genes are mobile elements that are embedded in intervening sequences, such as introns and inteins. Propagation of these genes occurs when the expressed homing endonuclease enzyme generates a double-stranded break at an intronless cognate allele, thus initiating homologous recombination pathways that repair the break via transfer of both the intron/intein and the homing endonuclease open reading frame (Figure 3) [12,21]. There are currently five known families of homing endonucleases; the largest and most thoroughly studied from a structural perspective is known as the LAGLIDADG endonucleases, named for the conserved amino acid sequence of a helix that contains the catalytic residues. These enzymes are usually encoded in the genomes of archaea, algal chloroplasts, and

the fungi mitochondria and are very specific, with target sites typically ranging from 18 to 22 base-pairs in length [12].

Endonucleases must continually balance the pressures of specificity and fidelity of recognition in order to maintain activity in the presence of genetic drift as well as to avoid toxicity to their target host [12,22,23]. In the case of the LAGLIDADG enzymes the reconciliation of these competing goals is elegantly expressed in the proteins' molecular structure (Figure 4, 2QOJ [24]). A variable specificity across the interface, with less specific positions often corresponding with wobble positions in the codons of the targeted gene, is achieved through a curved β-sheet structure that faces the major groove of the DNA. Alternating residue direction is a natural characteristic of β-sheets, and thus only every other protein residue interacts with DNA resulting in many nucleotides that are not completely saturated with direct contacts. This structural feature allows the endonuclease to tolerate changes to individual nucleotides in the site (relaxed fidelity), while still maintaining a high overall level of specificity due to the length of the DNA substrate.

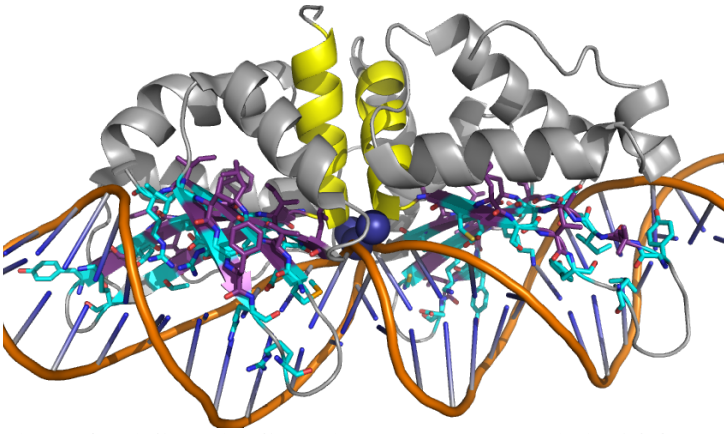


Figure 4. LAGLIDADG homing endonuclease I-AniI (2QOJ). Core residues (purple), DNA interacting residues (cyan), catalytic Mg^{2+} ions (blue), and helices each containing a LAGLIDADG motif (yellow).

Homing endonucleases have great potential as tools for targeted gene therapy and gene correction. These DNA cleavage enzymes are capable of inducing gene repair or disruption by stimulating DNA repair pathways at a specific location in a genome. Both computational design and directed evolution methods have been used to successfully altering the targeting specificity of homing endonucleases [11,25,26]. In my thesis research I expanded upon this work, improving both computational and experimental methods for altering endonuclease specificity, as well as completing detailed kinetics on wild-type and designed enzymes and incorporating information from homologues to guide interface engineering.

Chapter 2

Specificity Profiling of Wild-Type and Redesigned I-AniI Endonucleases[a]

Enzyme engineering should be informed by detailed analyses of the wild-type system. In my early work redesigning the target site specificity of the homing endonuclease I-AniI, it was discovered that, despite the approximate two-fold symmetry of the enzyme-DNA complex, there is almost complete segregation of the interactions responsible for substrate binding and transition state stabilization. This separation of the roles of these domains was revealed by doing kinetic studies on target sites differing by a single base-pair from the wild-type substrate. Computationally redesigned variants of I-AniI that achieved new specificities on one side did so by modulating binding, while redesigns with altered specificities on the other side modulated catalysis.

2.1 A Wild-Type Endonuclease Displays Distinct Binding and Cleavage Profiles

The model system used in many of the experiments presented in this work is the LAGLIDADG homing endonuclease I-AniI (Figure 4). While many of the well-characterized LAGLIDADG enzymes are homodimeric [22,27] I-AniI is a monomer, albeit with two distinct domains. One key advantage of working with a monomeric endonuclease is that engineering outcomes are more predictable because asymmetric functions and interactions in the sequence identical protein halves of a homodimer can be avoided [28]. The N-terminal domain of I-AniI makes extensive binding interactions to the left (-) side of the target site and the similarly structured C-terminal domain interacts with the right (+) side. This enzyme has been structurally characterized [24,29], improved with evolution [30], used as a model in the development of new yeast surface display technologies [31], and tested in the context of mammalian cell recombination assays [30,32].

[a] The following chapter is adapted with permission from the article "Exploitation of binding energy for catalysis and design" published in collaboration with Jordan Jarjour, Ryo Takeuchi, James J. Havranek, Justin Ashworth, Andrew M. Scharenberg, Barry L. Stoddard, and David Baker (Thyme, S. B. *et. al.*, *Nature* **461**, 1300-4, 2009).

I-AniI cleaves with high sequence specificity in the center of 20 base-pair DNA target site. Determining the nucleotide specificity at individual positions in this wild-type target site should precede engineering experiments that attempt to alter specificity. Endonucleases tolerate multiple nucleotides at some target site positions in order to maintain activity in the face of genetic drift and there is no reason to redesign an endonuclease to cleave base types that are already accepted by the wild-type interface, unless an increase in specificity in desired at these positions. Data of this sort is invaluable for improving computational modeling of protein-DNA interfaces. Benchmarking computational models with experimental data can guide improvements in the energy function and protocols. Computational recapitulation of specificity can also provide validation of homology models by giving an idea of the interface interactions in the absence of a crystal structure.

Control experiments probing the binding specificity of the I-AniI homing endonuclease, in preparation for computational redesign of specificity, revealed a striking asymmetry in the effect of base substitutions on binding affinity (Figure 5a). DNA cleavage and DNA binding by Y2 I-AniI endonuclease [30] were assayed for 60 different target sites, each containing a single base-pair substitution from the wild-type recognition sequence. Consistent with previous observations [24], enzyme activity assays showed that many nucleotide substitutions throughout the extended 20 base-pair recognition site abrogated or reduced cleavage, reflecting the high sequence specificity of the endonuclease (Figure 5b). Fluorescence binding experiments showed that for mutations between -10 and -3 on the (-) side of the interface, this loss of cleavage activity is associated with a loss of binding affinity. In sharp contrast, mutations in the -2 to +10 region of the recognition site, which also eliminated or reduced cleavage, had a minimal affect on substrate binding (Figure 5c).

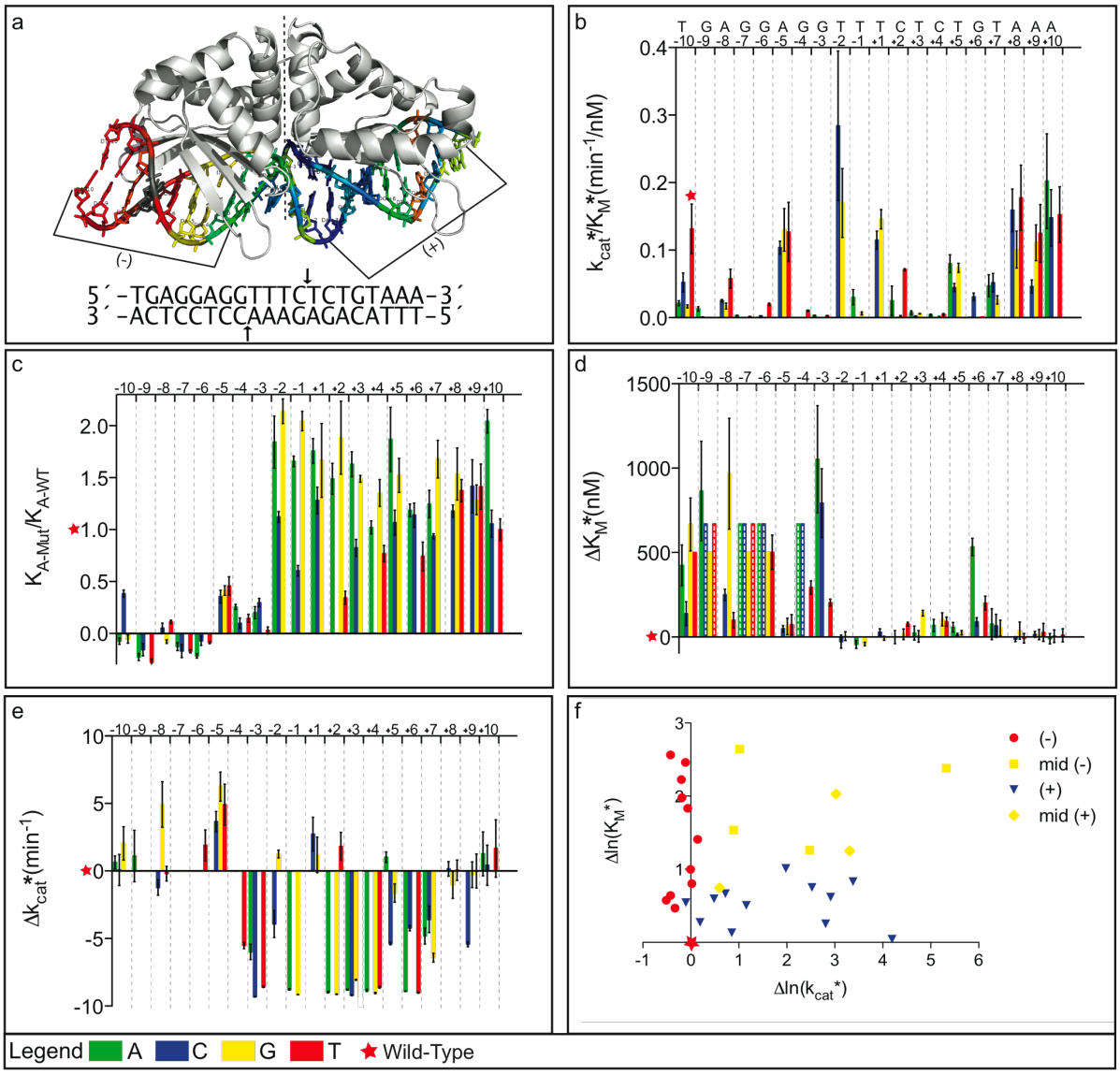


Figure 5. Segregation of contributions to binding and catalysis. a) Ribbon diagram of the I-AniI enzyme in complex with the wild-type target site (2QOJ). Target site and positions of DNA cleavage are shown below: (-) side cleavage site is cut prior to (+) side site. **b)** k_{cat}* / K_M* values for the wild-type target site (red star) and each of the 60 singly-substituted target sites (vertical bars). Substitutions throughout the length of the target site abrogate enzyme activity demonstrating the high sequence specificity of the enzyme. **c)** Relative binding affinities determined for each singly-substituted target site using fluorescence competition assays. Substitutions on the left side, but not the right side, significantly reduce binding affinity. **d)** K_M* values for each singly-substituted target site relative to the wild-type. As in c, substitutions on the left but not the right display significantly different values from wild-type. **e)** k_{cat}* values for each singly-substituted target site relative to the wild-type site. In contrast to c and d, substitutions between positions -4 and +9 have significant effects. Substitutions for which K_M* was too high (> 750nM) to allow separate determination of k_{cat}* and K_M* are indicated by bars with dashed lines in d, and are left blank in e. **f)** Asymmetry of the contributions to k_{cat}* and K_M*. Positions shown in red are on the left (-) side of the target site from -10 to -5 and almost exclusively contribute to K_M*. Positions shown in blue are on the right (+) side of the target site from positions +3 to +7. The boundary positions, -4, -3, and +6, contribute to both k_{cat}* and K_M* and are shown in yellow. To portray the structural context of these positions, the target site in a) is colored based on the effect of the mutation on k_{cat}*, normalized by the sum of the effects on k_{cat}* and K_M* ([|Δln(k_{cat}*)| / |(Δln(K_M*)|+ |Δln(k_{cat}*)|)] close to 1.0, blue; close to 0.0, red; intermediate, yellow; position where K_M* and k_{cat}* could not be separately determined; grey).

2.2 Kinetic Profiling Reveals Asymmetry in Interface Contributions to Catalysis

Enzymes utilize substrate binding energy both to promote ground state association and to selectively lower the energy of the reaction transition state [33]. Interactions between the enzyme and substrate promote catalysis both by bringing the substrate into close proximity and proper alignment with catalytic groups on the enzyme and by stabilizing the transition state for the chemical reaction [34,35,36]. Dissection of the contributions to enzyme catalysis has taken on renewed importance with the advent of computational and directed evolution approaches for engineering novel enzymatic activities for applications ranging from synthetic chemistry to therapeutics [37,38].

To determine whether the differences between the (-) and (+) side substitutions reflected differential contributions to ground state association versus transition state stabilization, the extent of cleavage of a linear double-stranded template as a function of time was determined for all 60 singly-substituted sites under single-turnover conditions, and pseudo-Michaelis-Menten parameters [39] K_M* and k_{cat}* were obtained from these data (Figure 6, Figure AI.1 and AI.2).

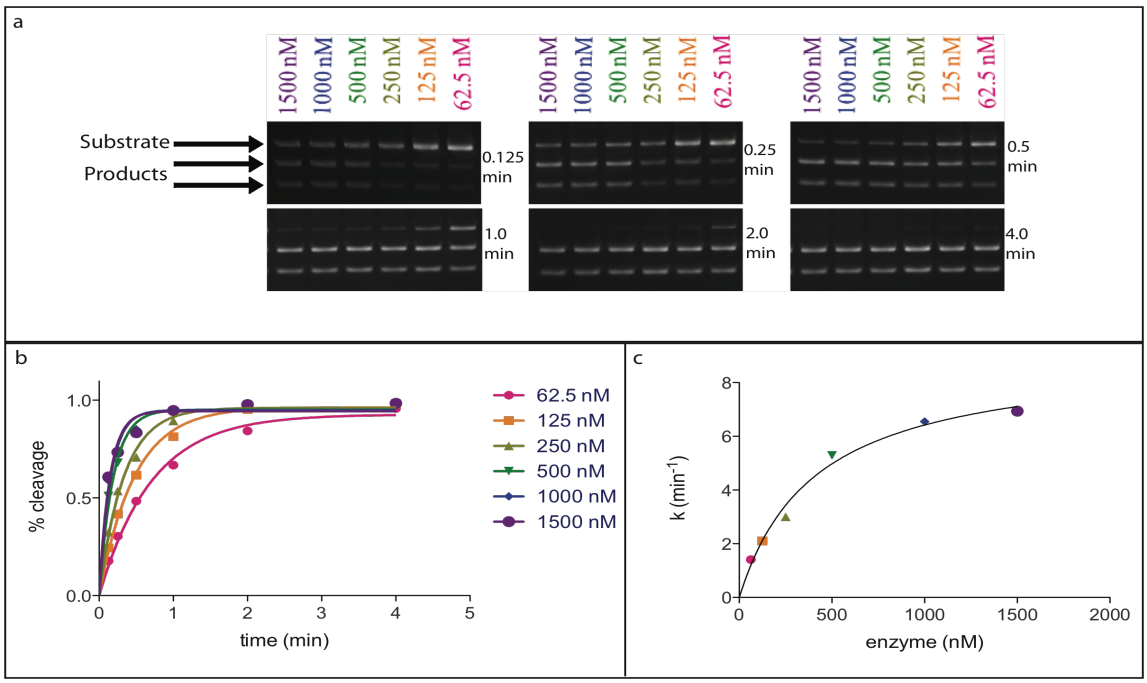


Figure 6. Example of a kinetic assay. Six time-points were collected over a 24-fold range of enzyme concentrations. b) The percent cleavage was calculated by integrating the densities of the bands in a) using ImageJ, and dividing the sum of the densities of the two product bands by the sum of the densities of all three bands. The percent cleavage for each enzyme concentration was plotted versus time using GraphPad Prism, and the rate was determined by fitting to a single-exponential function. c) The rate from each of the time-courses in b plotted versus enzyme concentration with Michaelis-Menten fits to determine the parameters k_{cat}* and K_M*.

Comparison of the k_{cat}* / K_M* for related substrates highlights the high sequence specificity of the enzyme: for example, at position -4 the k_{cat}* / K_M* for the wild-type G:C base-pair is >2000-fold greater than for A:T and >400-fold greater than for C:G (Fig. 5b and Table AI.1). The contribution of target site interactions to ground state stabilization (K_M*, Figure 5d) versus transition state stabilization (k_{cat}*, Figure 5e) was found to be skewed: substitutions on the (-) side increased K_M* significantly without reducing k_{cat}*, while substitutions on the (+) side decreased k_{cat}* with little effect on K_M*. The overall segregation of the kinetic contributions to specificity is shown graphically in Figure 5f and in the structural schematic in Figure 5a: most single base substitutions in the target affect k_{cat}* (blue, (+) side) or K_M* (red, (-) side) but not both. The striking feature of our results is that the apparent symmetry of the binding interface is completely broken during catalysis – chemically very similar protein-DNA contacts are utilized for substrate association on the left side and selective transition state stabilization on the right side.

Functional advantage of asymmetry

Our results suggest that initial binding of I-AniI to its target site involves formation of base-specific interactions on the (-) side and lower affinity non-specific interactions on the (+) side to form the Michaelis complex (the latter are suggested by yeast display experiments which show that the enzyme binds less tightly to the (-) half-site than to the full site [31]). Catalysis then requires bending of the DNA (note bend in Figure 5a), which is stabilized at the transition state by newly formed specific interactions between the (+) side and the enzyme. Such a two-stage mechanism (see following section) may be a general solution to the problem of specific target site recognition by enzymes that act on distorted DNA substrates. If the enzyme only bound to the distorted site, binding would require enzyme to be at the site (which may occur only once in the genome) simultaneous with fluctuation of the DNA into the distorted conformation; since both are rare events the net rate of binding, the product of two small numbers, would be very slow. If, instead, the enzyme can bind with some sequence specificity to undistorted target sites, the probability of being close enough to capture (and perhaps promote) fluctuations that distort the DNA will be very much higher. In I-AniI the total transition state binding energy appears to be roughly divided between the two steps: the N-terminal domain guides the enzyme to potential

target sites which match on the (-) side, and the C-terminal domain specifically stabilizes the transition state if there is also a match on the (+) side.

Two-stage DNA binding and domain dominance

The kinetic analyses presented in this paper suggests that in the ground state the interactions on the left side of the interface are largely intact, but those on the right side largely unformed. In the initial binding event, formation of the interactions on the left side may lead to a misalignment of the interactions on the right side, and a conformational change on either the DNA, enzyme, or both may be necessary to form the right side interactions. While a crystal structure of unbound I-AniI has not been determined, there such a structure of the monomeric endonuclease I-DmoI [40]; this structure is virtually identical to that in the complex with target site [41], suggesting that there is unlikely to be large structural changes in the enzyme during catalysis. In contrast, the target site in the I-AniI protein-DNA complex is bent considerably away from B-form conformation, and the enzyme displays a high enthalpic cost of protein-DNA binding which is likely associated with DNA deformation [42]. Yeast display experiments show that the enzyme binds more tightly to the (-) side than to the (+) side [31].

These findings are reminiscent of the concept of "domain dominance" described by Silva and coworkers [43] for I-DmoI. In this study it was found that C-terminal domain of the endonuclease loses significant specificity in the presence of Mn^{2+}, cleaving a palindromic target site containing a duplicated (-) half-site, while the N-terminal (A) domain maintains the same specificity in both Mn^{2+} and Mg^{2+}. Additionally, an engineered fusion of two N-terminal (A) domains of I-DmoI retains specificity throughout the entire target site in the presence of Mn^{2+}. Although the two subunits of I-AniI are optimized for different roles in DNA cleavage, the endonuclease maintains a very similar cleavage profile throughout the entire target site in both Mg^{2+} and Mn^{2+} (Michelle Scalley-Kim and Barry Stoddard, unpublished).

Additional monomeric LAGLIDADG endonucleases have also been demonstrated to display strong asymmetry in their binding affinity towards individual DNA half-sites, and/or significant difference in cleavage of those site's corresponding scissile phosphates. The yeast homing endonuclease I-SceI has higher affinity for binding to the 3' DNA half-site, leading to accumulation of nicked intermediates during the cleavage reaction [44]. Similarly, the algal

endonuclease I-CpaI preferentially nicks the bottom strand of its target site under limiting concentrations of metal ions [45]. Therefore, domain specialization may be a general property of monomeric LAGLIDADG homing endonucleases (as well as other DNA-binding proteins [46]), allowing rapid scanning and binding of B form DNA followed by specific stabilization of the kinked DNA conformations required for catalysis: an enzyme which bound only the bent conformations observed in the crystal structures would have binding kinetics limited by the population of rare DNA conformers, while one that bound only B form DNA would be incapable of catalysis.

Structural and computational comparisons with specificity data

The areas of interface interactions in I-AniI can be divided into categories based on their kinetic parameters. Sidechain-base-pair interactions from positions -10 to -5 are present in both the Michaelis complex and the transition state (base substitutions increase K_M* and K_D in solution and do not affect the rate when tethered). Sequence specific base-pair interactions from +3 to +8 are formed only in the transition state (substitutions have no effect on K_M* or K_D, reduce k_{cat}*, and slow the rate when tethered). A third class of interactions (at -5 and +7 for example) appear to be formed in the Michaelis complex but not the transition state (substitutions increase or decrease both k_{cat}* and K_M*/K_D).

Importantly for the design calculations described in the next section, three observations suggest that the crystal structure of the complex likely resembles the transition state more than the Michaelis complex: (1) specific interactions on the (+) side of the DNA target present in the crystal structure appear to be formed in the transition state but not the Michaelis complex, (2) the third class of substitutions mentioned above that appear to stabilize only the Michaelis complex make few interactions in the crystal structure (Figure 7), and (3) Rosetta specificity calculations based on the crystal structure correlate better with catalytic efficiency than with binding affinity (Figure 8).

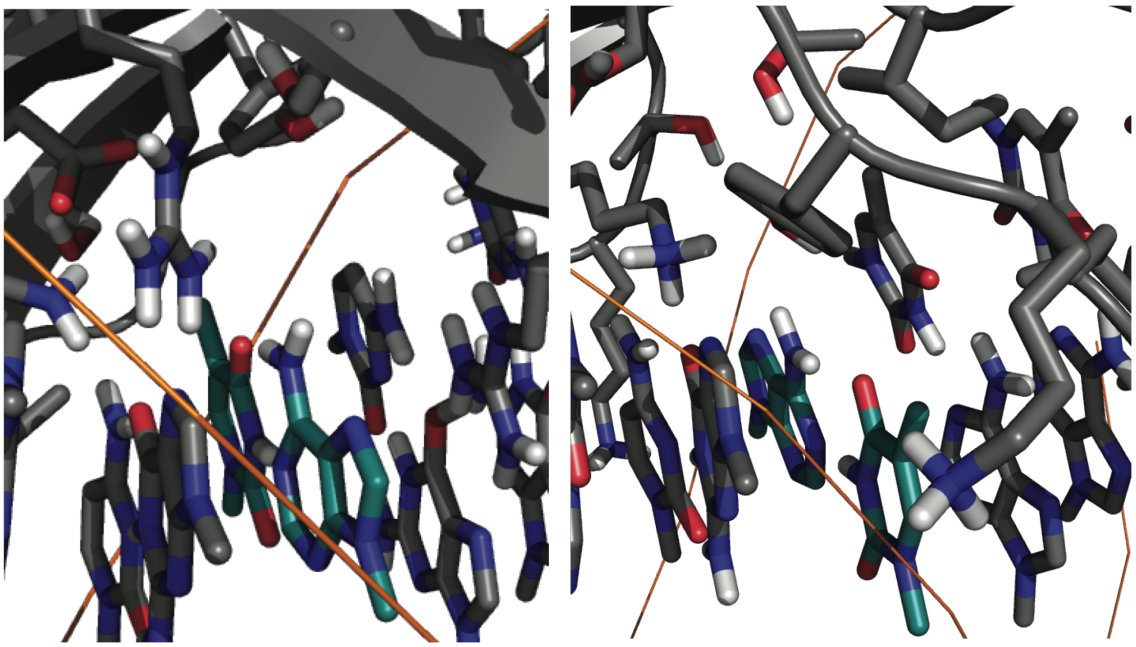


Figure 7. Positions predicted to make more interactions in Michaelis complex. The -5 position (left panel) and +7 position (right panel) do not make obvious contacts in the crystal structure. The -5 position has both an increased k_{cat}* and K_M* / K_A, and the +7 (as well as several other positions) has both a decreased k_{cat}* and K_A (K_M* values smaller than wild-type are difficult to determine accurately). This is consistent with removal or creation of interactions that are present in the Michaelis complex but not the transition state complex, and we predict these base-pairs make extensive interactions in the Michaelis complex than in the transition state complex.

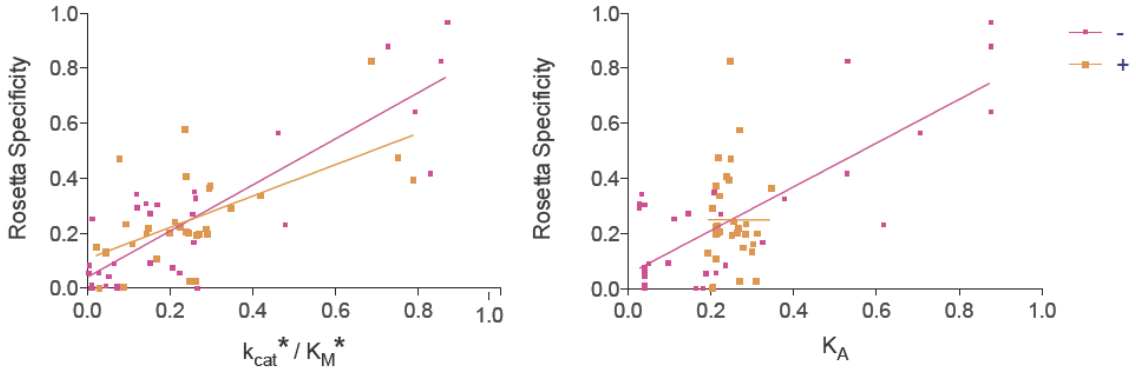


Figure 8. Correlation between predicted and observed specificity. k_{cat}* / K_M* is in left panel and K_A is in right panel. Specificity is defined as $e^{(-\Delta Xwt/t)} / \sum e^{(-\Delta Xall/t)}$ (where the numerator is the single base-pair of interest and the denominator is the sum of all four base-pairs at the position); t is set to 1.5. For the Rosetta computed specificities, ΔX in the numerator is the protein-DNA interaction energy, and in the denominator the sum is over all four basepairs at the position. For the experimental specificity determinations, ΔX is defined as $\Delta\ln(k_{cat}* / K_M*)$ (left panel) or $\Delta\ln(K_A)$ (right panel) for each substitution compared to wild-type. The correlation is significantly stronger for k_{cat}* / K_M* with Rosetta specificity (R^2=0.74 for left (-) side and 0.38 for right (+) side) than for K_A with Rosetta specificity (R^2=0.66 for left (-) side and 0.00 for right (+) side), supporting the idea that the crystal structure more closely resembles the transition state of the reaction than it does the ground state complex.

2.3 Kinetics of Redesigned Enzymes are Similar to the Wild-Type Enzyme

Monomeric LAGLIDADG homing endonucleases, which recognize non-palindromic targets, are attractive scaffolds for genome engineering applications [14]. An important challenge is to reprogram the substrate specificity of these enzymes towards desired target sequences. Rosetta computational design methods were used to redesign I-AniI to cleave sites with single base-pair substitutions. To redesign I-AniI specificity, the target site in the crystal structure of the I-AniI protein-DNA complex is mutated *in silico* and the program searches for combinations of amino acid substitutions that allow the formation of energetically favorable interactions with the new base-pairs, but not with the wild type base-pairs [25]. Design calculations were carried out for six target site variants bearing single base pair substitutions, genes encoding the amino acid sequences of eight redesigned enzymes were constructed, and the enzymes were purified. DNA cleavage assays revealed that the designed specificity changes were for the most part achieved (Figure 9, Figure AI.3, and Table AI.2). Our results demonstrate that I-AniI cleavage specificity can be reprogrammed by computational protein design, thereby providing starting points for the larger scale specificity changes required to cleave physiological target sites. In light of the interesting kinetics of this particular scaffold, a thorough kinetic analysis was carried out on eight designed endonucleases against their designed target sites and each of the three sites with an alternate base-pair at the redesign was completed.

Computational redesign methods for protein-DNA interactions

Computational protein design has over the last decade clearly been established as a leading player among protein engineering technologies, and one of the best-known programs for structure prediction and design is Rosetta, the program pioneered by the Baker group and under continual development by a worldwide community. There are several protocols for modification and analysis of protein-DNA interfaces using Rosetta [25,47]. Experimental success has been demonstrated with a protocol that uses Monte Carlo sampling to search identities and conformations (Dunbrack library derived rotamers [48]) of amino acids and a physically realistic all atom force field [25,49], with a modified version of that protocol that incorporates flexible backbone sampling with a cyclic coordinate descent (CCD) algorithm [50,51,52], and with a

genetic algorithm approach that simultaneously optimizes the protein sequence for binding affinity and specificity [50]. All of these methods rely on the physically based force field to score the interface and thus guide the redesign of interactions. A fourth protocol utilizes the geometries of protein-DNA interactions, or motifs, from published crystal structures to guide sampling [47]. While this motif-based methodology has not yet been experimentally validated, it is a promising approach that produces models with physically realistic interactions. One goal of my thesis work was to augment current protocols with motif derived interactions as well as to develop new protocols grounded in motif-based methodology. These advancements to the computational design methods are described in chapter 3.

Kinetics of redesigned I-AniI variants

An enzyme redesigned for a new target site could achieve altered specificity either by changing k_{cat}*, changing K_M* or changing both. To determine whether the designed changes in specificity were a result of changes in K_M* or k_{cat}*, for each of eight designed endonucleases we measured the single-turnover cleavage kinetics for target substrates containing each of the four possible base-pairs at the redesign position (Table AI.2). The details of three representative designs are shown in Figure 9. A design aimed at specific recognition of a DNA target site containing base-pair -8G:C (Figure 9a) achieved specificity exclusively by modulating K_M*: the K_M* decreased for the G:C, and increased for the A:T, T:A, and C:G. In contrast, a design aimed at specific recognition of +8C:G (Figure 9b) achieved specificity entirely through k_{cat}*: k_{cat}* decreased for A:T, G:C, and T:A, but was unchanged for +8C:G. Both of these designed enzymes have high specificity at neighboring base-pairs, and overall specificities that are higher than the wild-type enzyme in the targeted regions (Figure 10). A design aimed at specific recognition of the -3C:G substitution (Figure 9c), at the boundary between K_M* and k_{cat}* influencing positions (Figure 5), displayed changes in both k_{cat}* and K_M*, consistent with the results with the wild type enzyme at this position. These trends hold for the remaining designs as well (Table AI.2 and Figure AI.3): we find generally that the left side designs achieve specificity primarily by modulating ground state binding affinity, while the right side designs achieve specificity by modulating the free energy of the transition state.

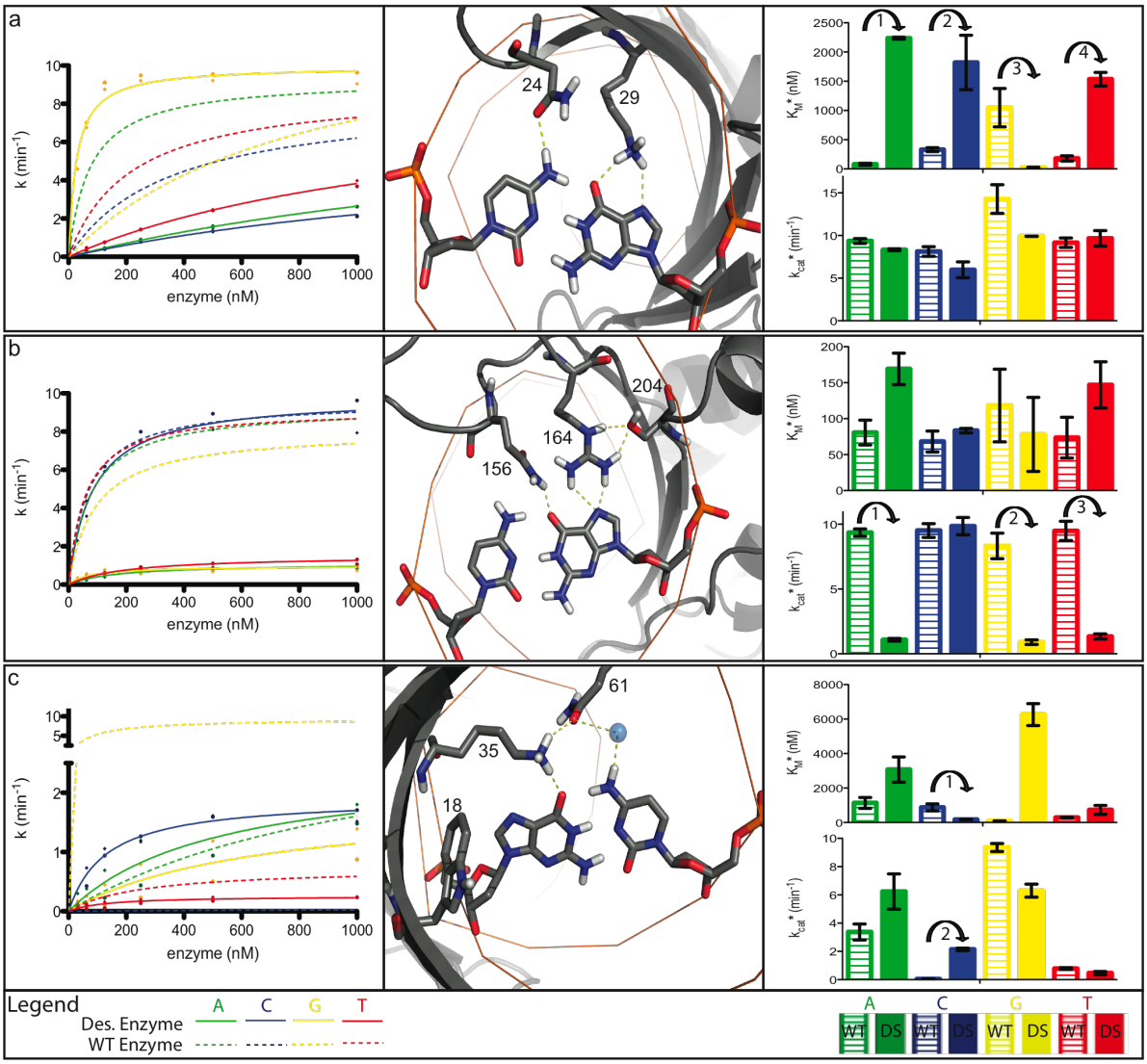


Figure 9. Computational redesign of specificity. Colour scheme: A, green; C, blue; G, yellow; T, red; error bars in right panels, s.e.m. a, Design for -8A:T to -8G:C substitution (K24N, T29K). Middle panel: the designed residues N24 and K29 make direct hydrogen bonds to -8G and -8C, respectively. Left panel: the concentration dependence of the cleavage activity for the designed enzyme (solid lines) for different base pairs at the -8 position differs considerably from the wild-type enzyme (dashed lines). Right panel: the k_{cat}* values remain approximately the same for both the wild-type and designed enzymes against all target sites, but the K_M* values are decreased for the target G base pair (arrow 3) and increased significantly for the other three substitutions (arrows 1, 2 and 4). b, Design for +8A:T to +8C:G substitution (L156Q, I164R, T204S). Middle panel: designed residues R164 and Q156 make direct hydrogen bonds to +8G. Designed residue S204 holds R164 in position. The kinetic traces (left panel) and bar graphs (right panel) show this design achieves altered specificity through changing k_{cat}*. The K_M* values remain approximately the same for both the wild-type and designed enzymes against all target sites, but the k_{cat}* values are significantly decreased for all of the competitor target sites (arrows 1, 2 and 3). c, Design for -3G:C to -3C:G substitution (Y18W, E35K, R61Q). Middle panel: designed residues K35 and Q61 make a direct hydrogen bond to -3G and a water-mediated hydrogen bond to -3C, respectively. Q61 and K35 also hydrogen bond with each other, and designed residue W18 further helps position K35 through packing interactions. The kinetic traces (left panel) and bar graphs (right panel) show this design achieves altered specificity through changing both k_{cat}* and K_M*. The designed enzyme has an increased k_{cat}* (arrow 2) and decreased K_M* for the -3C (arrow 1).

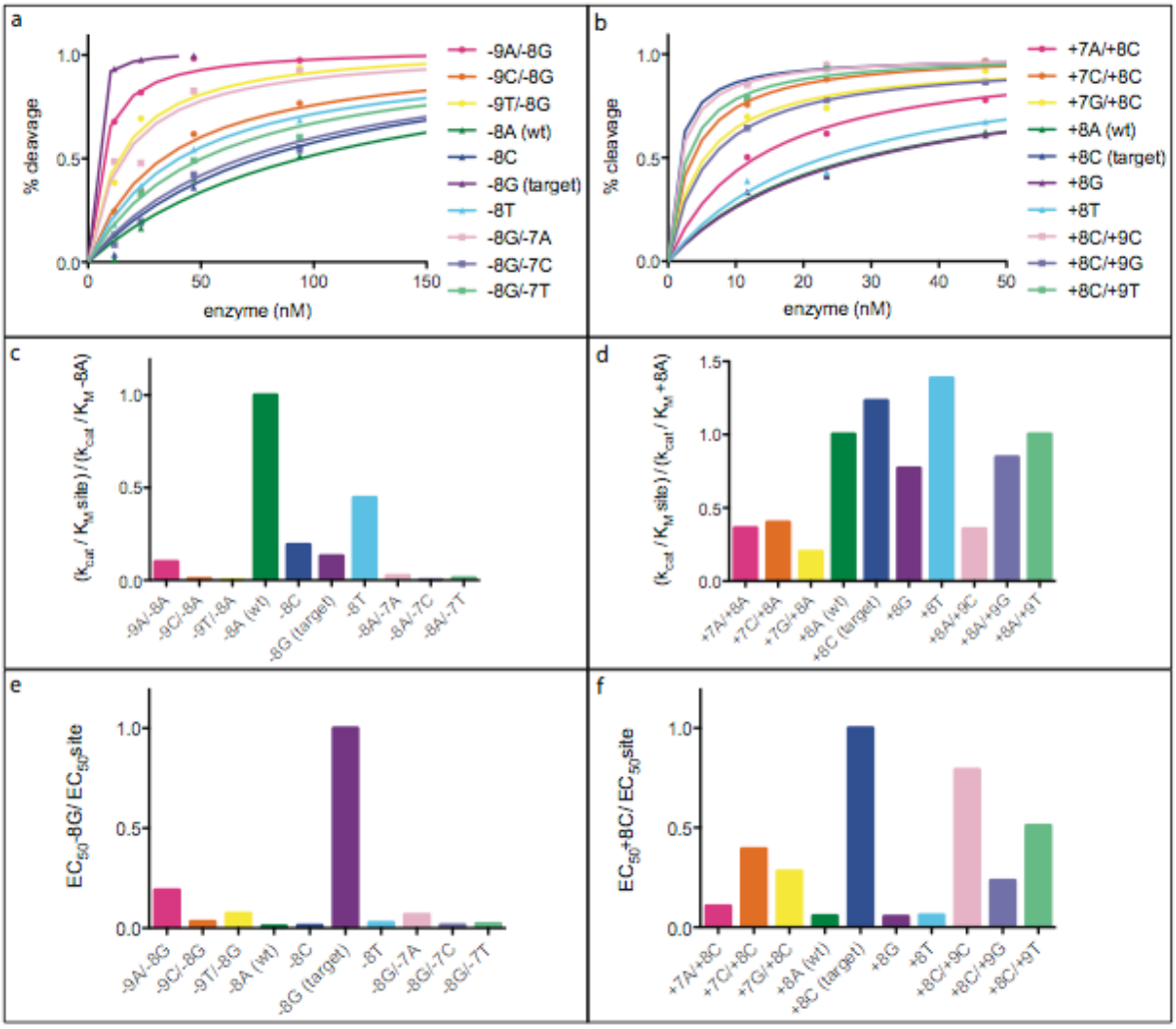


Figure 10. Adjacent specificities for two designs. Specificity at positions adjacent to the designed base-pair for the -8G:C_A and +8C:G designs. a) and b): percent cleavage by the -8G:C_A (a) and +8C:G (b) designed enzymes as a function of enzyme concentration. c) and d): specificity of Y2 (c) and designed enzyme (e) at positions surrounding -8G:C_A; d) and f): specificity of Y2 (d) and designed enzyme (f) at positions neighboring +8C:G. Specificity for Y2 (c) and (d) is determined by ratios of k_{cat}* / K_M* (Table S1). Specificity for designed enzymes -8G:C_A (e) and +8C:G is determined by ratios of EC_{50} values from (a) and (b). Comparison of c) and d) with e) and f) shows that the overall specificity of both designed endonucleases is increased over the Y2 starting point.

Implications for enzyme engineering

There is considerable synergy between classical enzymology and modern computational design. Design should be informed by detailed analyses of the wild-type enzyme since, depending on the enzyme and substrate concentrations in the application the designed enzymes are to be used for, it may be necessary to reengineer K_M*, k_{cat}*, and/or k_{cat}* / K_M*. Conversely, computational

design can provide insight into the basis for transition state stabilization. The union of classical enzymology with modern computational design, as illustrated here, provides a powerful approach to revealing the mechanistic basis for, and subsequently reprogramming, sequence dependent molecular recognition.

Chapter 3

Improving Modeling of Sidechain-Base Interactions[b]

Combinatorial sequence optimization for protein design requires libraries of discrete sidechain conformations. The discreteness of these libraries is problematic, particularly for long, polar sidechains, since favorable interactions can be missed. Previously, an approach to loop remodeling was described where protein backbone movement is directed by sidechain rotamers predicted to form interactions previously observed in native complexes (termed "motifs") [47]. The structural information in these motifs was incorporated into the Rosetta combinatorial sequence optimization protocols used for protein design and shown to improve native complex recapitulation. In this new implementation the motifs are used to bias both sampling and energetics of amino acid rotameric states in the context of a fixed protein backbone. Guided by the motif rotamer searches, the underlying Rosetta energy function was improved, increasing recapitulation of native interactions.

3.1 Using Structural Data to Improve Performance in Benchmarks

Advances in structural modeling algorithms for protein-DNA complexes lay the groundwork for functional predictions of these classes of interactions and engineering efforts. For example, accurate determination of binding specificity preferences for native complexes [53,54] and estimations of the contributions of individual amino acids to the energetics of an interface [55] can promote a better understanding of protein-DNA complexes and facilitate the next step: the computational refactoring of these properties for the development of tools for numerous biotechnology applications [14,56]. Improved computational methods have the capability to address the limitations of sampling size and significant experimental effort that constrain traditional combinatorial screening approaches [30,31,57] for engineering novel protein-DNA

[b] The following chapter is adapted with permission from the article "Improved Modeling of Side-chain-Base Interactions and Plasticity in Protein-DNA Interface Design" published in collaboration with David Baker and Philip Bradley (Thyme, S. B. *et. al.*, *J. Mol. Biol.* **419**, 255-74, 2012).

interactions. Currently, the main focus of protein-DNA interface engineering efforts is the reprogramming of DNA substrate specificity to alter binding or cleavage locations in a genome [25]. While there are a number of diverse experimental protocols to accomplish this engineering goal [30,31], the utilization of computational methods has been shown to complement and improve the efficiency of the experimental methods by guiding library design or providing a starting place for directed evolution [37,58,59].

The Rosetta macromolecular modeling and design suite [60] has been used for developing homing endonucleases with novel specificities [25,50,61,62]. Rosetta depends on a physically based energy function working in conjunction with a simulated annealing sampling algorithm to identify mutations in a protein that are likely to drive the formation of favorable, sequence-specific, protein-DNA interactions [49]. The general method for protein design with a fixed protein and DNA backbone involves a search of protein sequence and rotameric space to identify the predicted lowest-energy set of amino acid identities and conformations. Redesign for a specific DNA sequence change consists of substitution of the nucleotide type in the crystal structure DNA followed by redesign and repacking (search of rotameric, but not sequence space) of the amino acids surrounding this nucleotide change. A recent improvement to the Rosetta modeling of protein-DNA interactions was the incorporation of backbone flexibility on both sides of the interface, improving specificity predictions [53]. Backbone flexibility provides a way to further diversify design results over the standard, fixed-backbone approximation available in release versions of Rosetta. While the use of Rosetta has resulted in a number of endonucleases with successfully altered specificities [25,50,61,62], consistent recapitulation of experimental data has proven challenging [50,62], suggesting that many potentially successful designs are being overlooked by current algorithms.

The RCSB protein databank [63] contains within it a wealth of information in the form of the distances and geometries of protein-DNA interactions (“motifs”) present in native complexes (Figure 11). This information was incorporated into the Rosetta design process. Previously, motifs had been used to direct protein backbone sampling [47,64] and in this new implementation they are used to bias both sampling and energetics of amino acid rotameric states in the context of a fixed protein backbone. A library of these canonical amino acid-base interactions (motifs) was collected from protein-DNA complexes available in the protein databank (Figure 11). Rotameric conformations of amino acid sidechains capable of forming

interactions seen in that motif library were identified through a newly developed search process (Figure 12). This process scores the rotamers based on the distance between a canonical base placed in the motif-forming location and the closest base of the same type in the crystal structure. The rotamers that can form motif interactions, identified by a small distance between the canonical base and crystal structure base, are added, with an energetic bonus, to the rotamer set

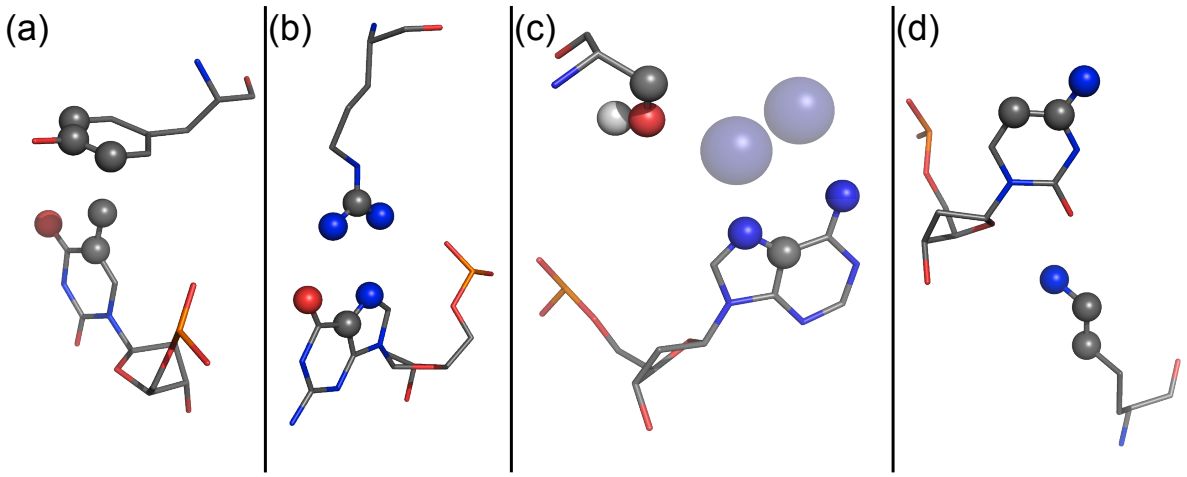


Figure 11. Examples of the types of motif interactions. Atoms that define the motif interaction are shown as spheres colored by atom type. (a) Tyrosine residue packing against a thymine methyl group, derived from Tyr25A and Thy317B of 1mow. (b) Bidentate arginine-guanine interaction, derived from Arg274B and Gua418C of 1cyq. (c) Water-mediated interaction identified by placement of waters (transparent blue spheres) on the DNA at canonical locations, derived from Ser47A and Ade516C of 1m5x. (d) Minor groove interaction, derived from Lys116A and Cyt16C of 2np6.

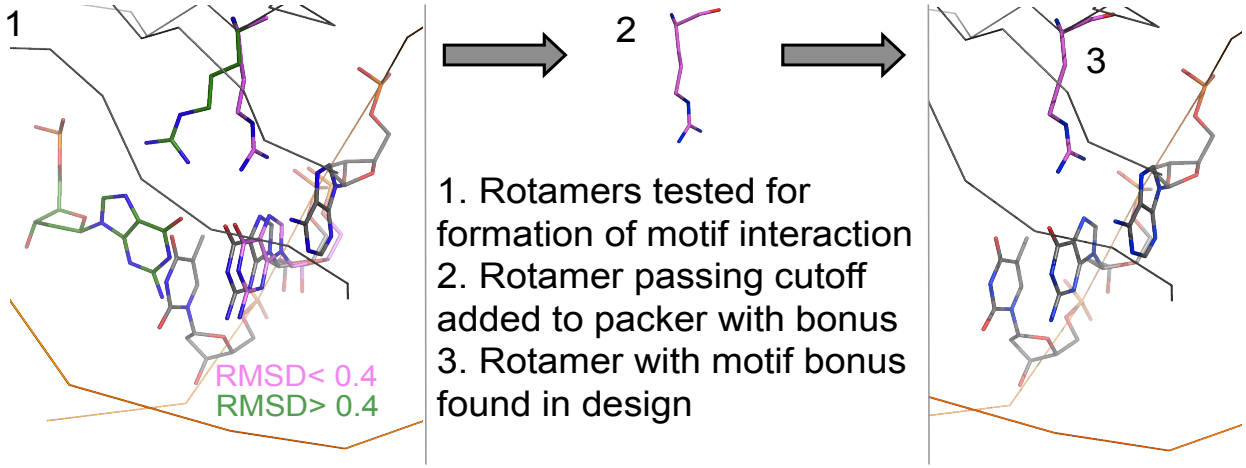


Figure 12. Overview of the motif-biased design protocol. In step 1, a series of rotamers and motifs are tested to see if they are compatible with the crystal structure undergoing design. These rotamers and motifs are subject to a series of cutoffs: distance of C1*, how parallel the placed base is to the crystal structure DNA, and RMSD of nucleobase atoms. In this example, two arginine rotamers (green and pink) are tested with a bidentate arginine-guanine motif and the pink rotamer passes a nucleobase RMSD cutoff of <0.4 when an ideal guanine base is placed in a motif-compatible position and compared to the nearest guanine base in the crystal structure. This pink arginine rotamer is then added to the standard rotamer sets used by the Rosetta packer. The rotamer is given an energy bonus over other rotamers and is found in a design completed for this guanine base.

used by the standard, fixed-backbone Rosetta design protocol. The size of the rotamer library used in standard design calculations is limited due to computational considerations, and this search process allows assessment of many more rotamers than could normally be included. While only a small fraction of the screened rotamers are added to the rotamer library – the procedure is limited to 100 extra rotamers of each amino acid type at each position – the incorporation of these interaction-biased sidechains provides a way to increase exploration in areas of sequence and rotameric space that are most likely to result in the formation of native-like contacts.

In order to analyze the effect on design of adding these motif-biased rotamers and determine the optimal bonus value for them, calculations were carried out for a set of 112 protein-DNA co-crystal structures. This set was divided into a training set of 48 proteins and a test set of 64 proteins for assessing the validity of protocol optimizations found to improve results for the training set. The sequence recovery for this test set, analyzed by two metrics ("weighted" and "unweighted" recovery), is shown in Figure 13a for a range of motif bonus values. The addition of motif rotamers was found to improve the sequence recovery for both recovery metrics, across multiple variants of the Rosetta energy function (Figure 13a). Examining sequence recovery as a function of the motif bonus term revealed that low bonuses generally give the best results. Values of -1.25 or -2.50 Rosetta Energy Units (REUs – most closely correlated with kcal/mol [65]), depending on the other scoring parameters and the recovery metric, resulted in optimal recovery. Higher bonus values have reduced recovery due to the incorporation of motif rotamers without regard to other energy function terms. The motif bonus resulting in the highest sequence recovery was slightly less for the weighted metric than for the unweighted metric. The unweighted metric counts every designed position equally, and is thus subject to a bias favoring incorporation of the amino acid types most commonly found in protein-DNA interfaces (such as those types in the motif library). The weighted metric is an average over the recoveries for each amino acid type and free from biases in the amino acid composition of the interface positions. Accordingly, the very high

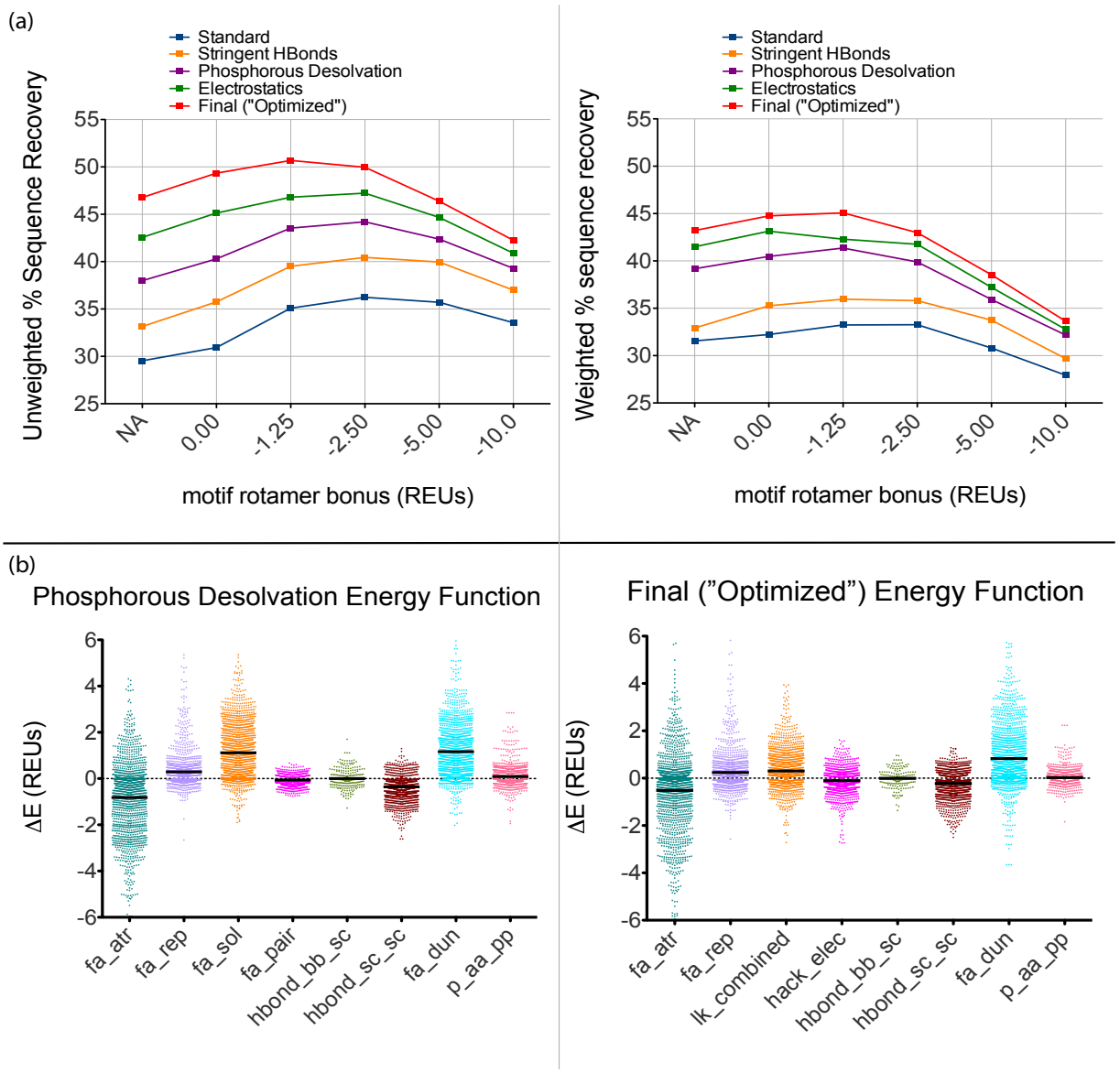


Figure 13. Optimization of Rosetta energy function. Abbreviations for energy function terms: fa_atr = attractive, fa_rep = repulsive, fa_sol = solvation, fa_pair = distance-dependent atom pair potential, hbond_bb_sc = hydrogen bonds between backbone and sidechain atoms, hbond_sc_sc = hydrogen bonds between sidechain atoms, fa_dun = rotamer probability, p_aa_pp = probability of amino acid given backbone conformation, hack_elec = simple electrostatics, lk_combined = combination of terms for orientation-dependent desolvation model. (a) A comparison to two metrics of sequence recovery over several motif rotamer bonuses and several iterations of energy function optimization (Figures AI.5 – AI.8). The "Standard" energy function was the starting point for the optimization. The "Standard" energy function was improved by the addition of motifs, increasing the stringency of the hydrogen-bonding model ("Stringent HBonds"), modification of the phosphorous desolvation penalty ("Phos. Desolvation"), and the addition of a coulombic electrostatics term1 for the "Electrostatics" energy function. The "Final ("Optimized")" energy function includes multiple additional changes detailed further in the text and in the supplement. (b) Energy differences, separated out by energy term, between incorrectly designed rotamers and rotamers with a motif bonus that match the native amino acid type, or more correctly match the native rotamer, than a designed rotamer with no bonus. The units for these energy differences are in Rosetta Energy Units or REUs. The differences collected with the "Standard" energy function reveal that the solvation term (fa_sol) and the rotamer probability term [48] (fa_dun) are the two energy terms that are being offset by the motif bonus. As a part of the energy function optimization the solvation term was replaced with an orientation-dependent solvation model [53] (lk_combined) and changes were made to the atom-specific desolvation parameters for several amino acid types.

motif bonus values were less detrimental to unweighted recovery, which benefited from biases toward abundant amino acid types, than to the weighted metric.

3.2 Optimization of the Rosetta Energy Function

We next used the motif-biased design results to guide optimization of the Rosetta energy function, improving sequence recovery significantly over "Standard" scoring. The complete set of modifications to the energy function resulted in a high unweighted recovery of 50.7% with motifs added, an increase of 20% over the initial "Standard" recovery of 29.6% with no motif rotamers or optimization (Figure 13a, Table AI.3). The recovery pattern and the magnitude of the differences in recovery observed for this test set are similar to those changes seen for the training set, over the same iterations of the energy function (Figure AI.4).

Of these scoring improvements, many were implemented specifically for modeling of protein-DNA interactions, such as increase in the stringency of the hydrogen bonding model and correction of the Rosetta phosphorous desolvation [66] parameter (Figure 13a) [53]. The combination of this corrected solvation model and the increased hydrogen bond stringency provides over 8% of the total 20% improvement in unweighted recovery. The change having the next largest effect was the replacement of the database-derived, residue-pair potential (the fa_pair term) with a simple, short-range explicit electrostatics term [53]. Recoveries with only this "Electrostatics" modification are shown in Figure 3a. Both the electrostatics model and the motif bonus favor charged interactions – charged residues are overrepresented in the motif library due to their abundance at protein-DNA interfaces – thus a higher motif weight is less beneficial in the presence of the electrostatics model (Figure 13a, comparing "Phosphorous Desolvation" to "Electrostatics"). The "Final" optimized scoring function garners further improvements in recovery of over 4% unweighted (1.7% weighted). This finalized scoring function is a composite of several smaller improvements the individual effects of which are detailed in the supplement (Figure AI.5-AI.8). These changes are 1) a modification to the solvation model (lk_ball), introduced by Yanover and Bradley [53], in which desolvation contributions for polar atoms are dependent on the relative orientation of the desolvating atom, 2) the modification of desolvation parameters for atom types found in asparagine, glutamine, lysine, and arginine amino acids, 3) an increased weight of the attractive (fa_atr) scoring term, 4)

an increased positive charge for the lysine NH3 group as a proxy for an inability in Rosetta to differentially weight hydrogen bonding types, and 5) an optimization of the amino acid specific reference energies.

This optimization of the Rosetta energy function was guided in part by analyzing the biases in the sequence recovery results. Examining the ratio of the number of times an amino acid was designed to the number of times it is found in the initial population reveals amino acid types that are under- and over-represented by the design process. All modifications to the desolvation terms, as well as the increased positive charge of lysine, were prompted by a low recovery of those amino acid types and a corresponding low representation of these types in the designs completed using the energy function with only the electrostatics term added. The sequence recoveries and amino acid ratios leading to and resulting from each modification are detailed in Figure AI.5-AI.8. Optimization of the amino acid specific reference energies, representing the average energy of the residue in the unfolded state, was also guided by looking for biases in the distribution of designed amino acids.

In addition to correcting biases in amino acid composition, a comparison between designs completed with and without motifs highlighted the energy terms most in need of optimization. The sequence recoveries of designs with a bonus on motifs were higher than those without the added motif rotamers. Determination of those energy terms that were offset by the motif bonus helped to guide our energy function optimization. If a motif rotamer of the native amino acid type is incorporated in a design and more closely matches the wild-type rotamer than an incorrectly designed rotamer without a motif bonus, the differences in energy terms between the motif rotamer and incorrect rotamer can illuminate what terms are responsible for favoring the incorrect rotamer. This analysis was completed over the entire set of 112 designed interfaces and the results for the “Phosphorous Desolvation” and “Final (“Optimized”)” weight sets are shown in Figure 13b. Energy differences with a positive value are the ones being offset by the motif bonus for the more correct rotamer choice. For the starting energy function the two energy terms that are positively shifted are the solvation (fa_sol) and rotamer probability [48] (fa_dun). The final energy function indicates that the design failures associated with a solvation penalty were significantly corrected by a combination of the modifications to desolvation terms and the addition of the orientation-dependent solvation model. Ways to correct the remaining penalty associated with the rotamer probability term are currently under study. These findings correlate

with the shift toward a preference for lower motif weights in concert with higher sequence recovery as the energy function was optimized. This result indicates that more successful motif-like interactions were being made without the aid of such significant motif favoring as energy function improvements were incorporated.

Chapter 4

Modeling Natural Plasticity in Protein-DNA Interfaces[c]

To further test the methods, we carried out a comprehensive experimental scan of amino acid preferences in the I-AniI protein-DNA interface, and found that many positions tolerated multiple amino acids. This sequence plasticity is not observed in the computational results because of the fixed backbone approximation of the model. We improved modeling of this diversity by introducing DNA flexibility and reducing the convergence of the simulated annealing algorithm that drives the design process. In this work, we developed methods for exploring energetically relevant sequence diversity in order to produce designs enriched in amino acids making native-like interactions with the DNA bases. These new methods are potentially valuable for guiding design of libraries for experimental engineering methods, and their success was evaluated by comparison to a newly collected experimental dataset. New protocols for increased diversity generation included differential energetic and sequence-space biasing for rotamers capable of forming canonical motif contacts, simulations with flexible DNA [53], and reducing the convergence of the simulated annealing algorithm. The resulting predictions were analyzed in the context of sequence recovery benchmarks and a newly generated comprehensive experimental dataset that identified the tolerated sequence variation at 44 positions in one protein-DNA interface. In addition to serving as a benchmark, this extensive experimental dataset provides insight into the types of interactions essential to maintain the function of this potential gene therapy reagent.

4.1 Sequence Optimality Screen Reveals Endonuclease Interface Plasticity

A designed amino acid that does not match the native sequence is not necessarily a failure of the computational methods. Depending on the physiological role of a DNA-binding protein, the wild-type amino acid may not be the most energetically favorable. Some regions of a protein-DNA interface may require low specificity and hence few direct nucleotide contacts in order to accommodate multiple DNA bases – such as transcription factors that must bind to multiple

[c] The following chapter is adapted with permission from the article "Improved Modeling of Side-chain-Base Interactions and Plasticity in Protein-DNA Interface Design" published in collaboration with David Baker and Philip Bradley (Thyme, S. B. *et. al.*, *J. Mol. Biol.* **419**, 255-74, 2012).

promoters [67]. While some protein positions in an interface require the wild-type amino acid for activity or binding, other positions can tolerate multiple amino acid types. Without knowing the role and importance of each amino acid in an interface, it is insufficient to use sequence recovery of native interfaces as the sole metric for determining the success of the computational methods. A straightforward way to address this question is to make and characterize protein mutations and see if they are tolerated or disallowed as computationally predicted. This experiment was carried out for one protein in the benchmark set, the homing endonuclease I-AniI. Full randomization of each of 44 positions in the interface of the homing endonuclease I-AniI and screening of all single-position libraries for activity against the wild-type target site was completed using a bacterial directed evolution system [26]. Sequencing ~20 protein mutants for each library (Table AI.4) after activity selection showed which positions tolerated only the wild-type amino acid and which positions could accept a number of amino acids.

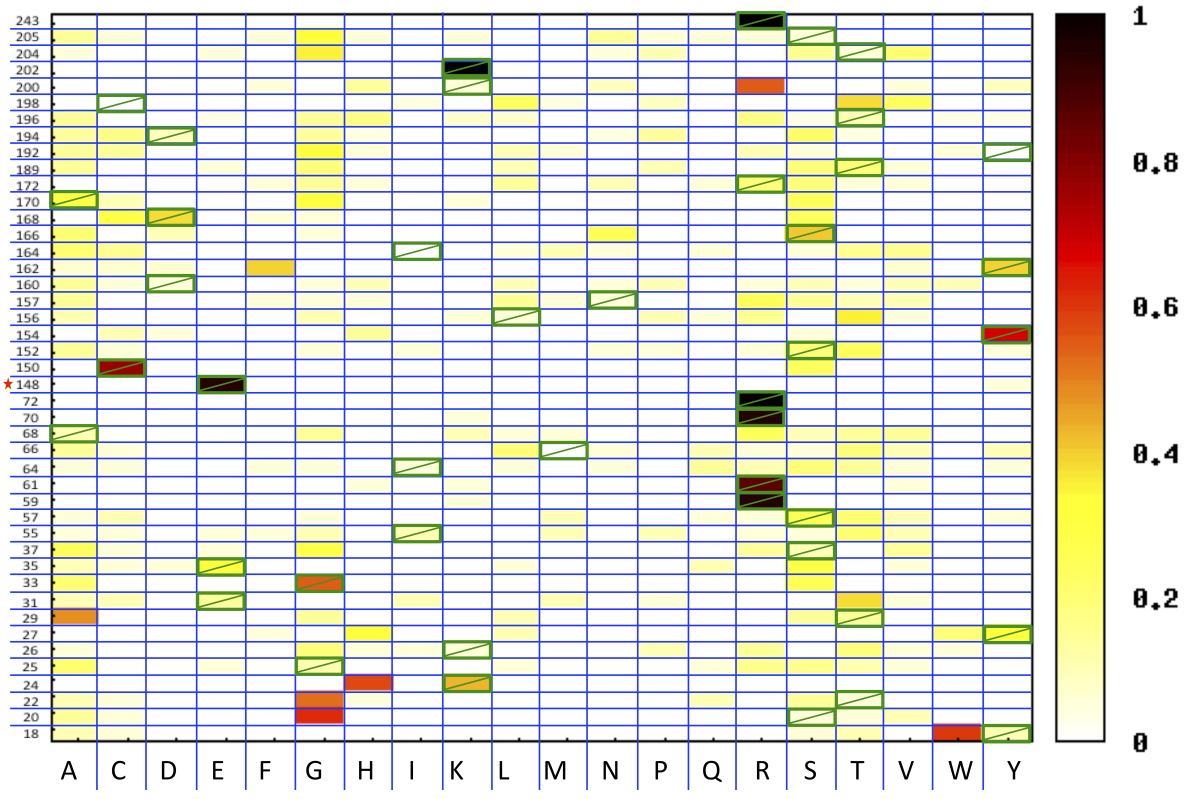


Figure 14. Sequence optimality of the interface residues of I-AniI. Heatmap displaying the frequencies observed of each amino acid type in a selected pool of sequences at each of 44 positions in the I-AniI interface. The wild-type amino acid is marked with a green box. Each position in the interface was fully randomized and these single-position libraries were subject to an activity selection [26]. A frequency of 1 means that the amino acid with this frequency was the only amino acid type observed at that protein position, whereas a frequency of 0.05 would be an amino acid type observed once from a set of 20 sequences.

The experimental data revealed that the wild-type amino acid type is not favored highly over other possibilities at many positions in the interface (Figure 14). The calculated experimental recovery, an average over all wild-type recovery frequencies, is 31%. Only a few positions show very high preservation of the wild-type amino acid. In the N-terminal domain, only four arginine residues are preserved, certainly contributing significant binding energy (R59, R61, R70, and R72). In the C-terminal domain, preserved residues include the position Arg243, stabilizing the position of a C-terminal DNA-contacting loop through interactions with the protein backbone, and interacting amino acids Lys202 and Tyr154, likely key contributors to formation of the catalytic complex [61]. The importance of these three C-terminal residues for cleavage of this particular target DNA is underscored by their complete conservation in homologues of I-AniI predicted to cleave a very similar target DNA sequence, even in those with sequence identity of less than 50% [68]. The other aromatic residue positions on both sides of the interface display higher conservation in this dataset than the majority of positions, with the exception of Tyr192. While these aromatics did not always show a high recovery of the exact

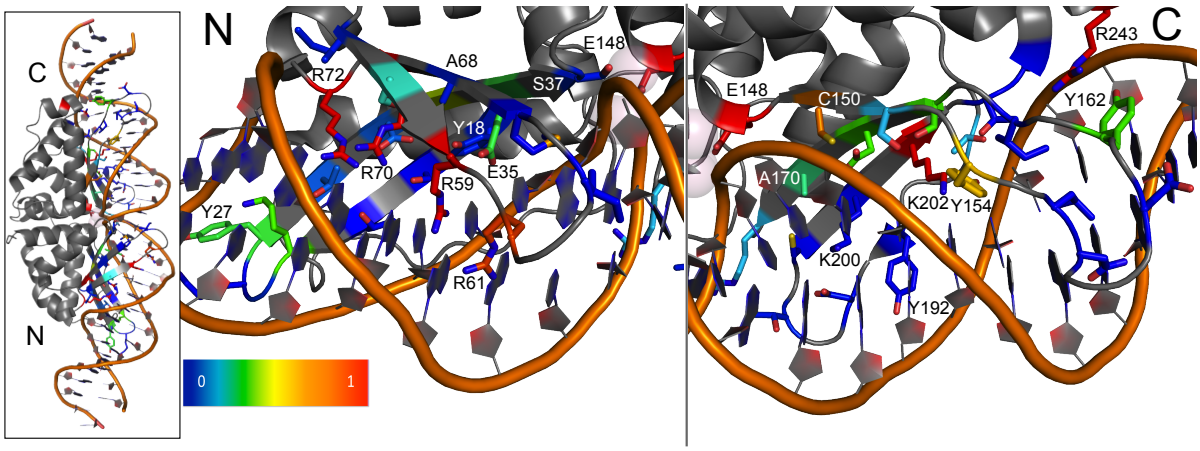


Figure 15. Visual representation of the interface conservation of I-AniI. The frequency of observing the wild-type amino acid after full randomization and selection (Figure 14) is summarized on the structure of I-AniI (2QOJ). Only the 44 residues that were randomized are shown in this representation. Blue corresponds to a frequency of zero, or non-conserved positions. Red corresponds to positions that are highly conserved as the wild-type amino acid. The overall protein-DNA complex is shown in the leftmost panel, and the N- and C-terminal domains are separated in the other panels to allow for a closer examination of the conserved contacts. Four arginine residues are most conserved in the N-terminal domain, and are likely essential for formation of the initial substrate-bound complex. Lys202 and Tyr154 are conserved in the C-terminal domain and these interactions likely play an important role in the formation of the catalytic complex [61]. This representation is incomplete in that it loses information if the preferred amino acid is not the wild-type, but still a conserved type. For example, positions Tyr18, Tyr27, and Tyr162 are strongly conserved as aromatic residues (Figure 14), but the native aromatic shows up at lower or equivalent frequencies as other aromatic types, resulting in blue or green shading at these positions.

native amino acid type, they all displayed a tendency to remain an aromatic. The frequency of recovering the wild-type amino acid at each position is visually presented on the I-AniI structure (2QOJ [24]) using a gradient from red to blue; positions that come back as wild-type are colored

red and the positions with very little wild-type observed in the sequencing results are blue (Figure 15).

The significant number of positions displaying little or no preference indicates that many amino acid substitutions in the I-AniI interface are functionally neutral, in least in the context of this selection system. The ability of the interface to accommodate such neutral drift – the accumulation of non-deleterious mutations with adaptive potential – has been implicated as a mechanism for the acquisition of new substrate specificities [68,69,70]. This neutral drift facilitates enzyme adaptations by reducing the number of mutations necessary to acquire new functions in the face of evolutionary pressure, and is particularly important for the endonuclease family of proteins. These DNA-cleaving enzymes are parasitic elements, catalyzing transfer of their own gene, and their interface flexibility allows for their continued propagation by facilitating cleavage of a wide range of target sites that are themselves subject to genetic drift.

Numerous positions show very low levels of wild-type amino acid in the sequencing results (at or below 5%, or 1 of 20 sequences) and understanding how differences in frequency correlate with differences in enzyme activity is important for utilizing this dataset. When there is strong selective pressure the position converged almost completely to the preferred sequence, such as in the case of the magnesium-binding catalytic residue Glu148 that was randomized as a control for the experiment (Figure 14, 15). This assay of activity is also sensitive to small differences in activity, as is demonstrated by the data collected for position Lys200. K200R and K200N were previously tested mutants, since they were both observed in homologues of I-AniI and shown to have levels of activity very similar to wild-type [68]. Both mutants were found to be slightly more active than the wild-type enzyme, and in this current assay they both were found in the selected pool with higher frequencies than the wild-type lysine (0.55 for Arg, 0.09 for Asn, and 0.05 for Lys). Given the extremely high activity of both mutants, it was challenging to resolve whether one was more active than the other with previously published enzymatic cleavage assays [68]. However, arginine was by far the most common amino acid observed at position 200 in an alignment of homologous enzymes [68] (Figure AI.9), matching the data here showing that it is observed more frequently than any other amino acid in the selected pool (Figure 14). While the amino acid frequencies at this particular position match those observed in a multiple sequence alignment of endonucleases predicted to cut a very similar site to I-AniI, the majority of the positions observed experimentally to have high flexibility are significantly less

variable in the alignment (Figure AI.9). The conditions of the bacterial selection system differ from natural evolution, likely resulting in this divergence between the alignment and the results observed from the described experiments. In particular, the bacterial system is selecting only for activity on the wild-type I-AniI, not for specificity against competing target sites or lack of specificity at areas facilitating new specificity acquisition, and artificial selections allow for full randomization at any interface position, whereas natural evolution generally traverses a pathway constrained by single nucleotide substitutions in the starting codon.

4.2 Computational Methods for Sequence Diversity Generation

The high sequence diversity tolerated at many positions in the I-AniI interface points to the need for computational protocols that generate multiple, energetically reasonable solutions, rather than a single design. Algorithms that produce only a lowest energy solution are constrained by sampling and the quality of the energy function guiding the design process. Methods are needed to generate diverse structures, thus enabling new local minima to be found. Diversity in design is valuable for comparison to experimental data, as library-screening experiments rarely produce a single best protein sequence for a given target and instead provide several solutions. Multiple low-energy solutions can also be screened concurrently in directed evolution experiments.

Two methods, DNA backbone flexibility and reducing the convergence of the simulated annealing algorithm ("the packer" [60]) used by the Rosetta, were developed and assessed in the context of a computational benchmark and experimental data. The DNA flexibility consisted of a 3 base-pair pocket of movement surrounding the target design base-pair (Figure 16a, "DNA-Rebuild"), and the convergence of the packer was reduced by increasing the low temperature of the simulated annealing procedure and removal of the quenching step that drives the packer to identify the sequence with the lowest possible energy ("HighTemp-Packer"). Out of the full set of 112 proteins, a complete set of interface designs was collected with both of these new protocols for 78 that were compatible with the DNA-Rebuild methods in their current state. All data was collected with the "Optimized" energy function. No motif rotamers were added for these computational experiments. A total of 56 designs were completed for every design pocket (DNA base-pair and surrounding protein positions) that was previously designed a single time with the standard design protocols. The frequencies of amino acids observed at each designable

position were calculated over these 56 designs and compared to frequencies from 56 designs completed with the standard method.

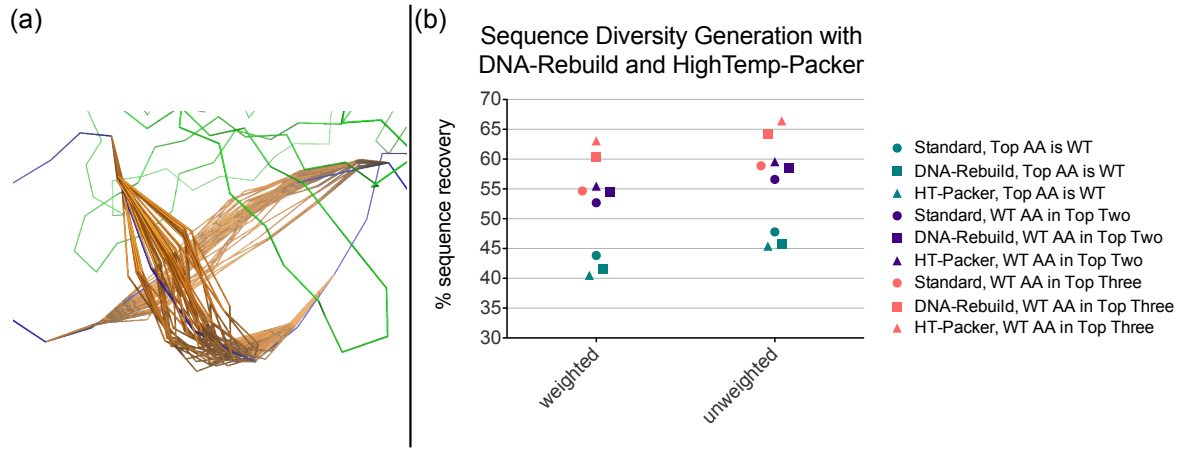


Figure 16. Limited degeneracy increases sampling of the native sequence. (a) Illustrative example of the level of DNA movement in the DNA-rebuilding simulations. (b) Both methods developed for sampling diverse sequences were tested, and compared to the "Standard" method, for a benchmark set of 78 proteins. The frequencies of amino acids observed at each position were calculated from 56 trajectories for each method. If only the highest frequency amino acid is incorporated in the sequence recovery calculation (cyan), the recovery shows a slight decrease for both weighted and unweighted metrics. If the top two (purple) or top three (pink) amino acids are both considered in the recovery calculation, and observing the wild-type amino acid in any of these top positions counts as correct, then the sequence recoveries are significantly increased.

The results of both protocols on the two sequence recovery metrics revealed that the diversity produced often contains the wild-type amino acid, even if it is not the most frequently observed type at a particular position. If the top two amino acids by frequency were considered when calculating recovery, the chance of correctly identifying the wild-type is increased over 12% for both recovery metrics (Figure 16b). However, while the sequence variation is much less for the 56 design runs with the standard protocol, recovery with this original method also improves by 8% when the top two amino acids are counted, achieving a high of only about 2% lower than the two new methods. Looking at the top three most frequent amino acids drastically increases the recovery gap between the original method and these new methods that generate significant sequence diversity. The HighTemp-Packer achieves a highest unweighted recovery of 66.4%, a 7% improvement over taking only the top two amino acids. The DNA-Rebuild performs slightly less well, achieving only 64.3% unweighted recovery, but still significantly outperforms the original method that only shows a 2% gain to 58.9% unweighted recovery. Computational results that produce possible amino acid choices, rather than a single lowest-energy choice are essential for building libraries to guide experimental engineering projects. However, the success of building libraries based on this expanded sequence pool requires that the added information increase the chance of finding a native-like or low-energy state rather than

simply diluting the good sequences with inaccurately produced diversity. The result that both of these new protocols significantly improved sequence recovery when the second or third highest frequency amino acids were added to the recovery calculation argues that both protocols could add valuable diversity to a designed library. Comparisons to experimental data conducted in the next section further explore the merits and limitations of both methods.

4.3 Computational Recapitulation of Experimental Sequence Optimality Dataset

Comparison of the experimental data with the previously described computational protocols indicates that neither of the new protocols stands out as superior and that each method has different strengths (Figure 17, Figure AI.10-AI.11). Both protocols better recapitulate the experimental data than the "Standard" design method (Table 1 [71]).

Table 1. Comparison of computational protocols to experimental data. Divergence between experimentally observed and computationally predicted amino acid frequency distributions at 44 positions of the I-AniI protein-DNA interface was assessed using two standard metrics for comparing probability distributions: the Jensen-Shannon divergence [71] and the Euclidean distance. A lower divergence value indicates that the probability distributions better match one another.

Computational Method	Jensen-Shannon divergence	Euclidean distance
Standard	0.472	0.839
DNA-Rebuild	0.409	0.670
HighTemp-Packer	0.399	0.695

The amino acid frequencies observed at some positions better matched the frequencies from the DNA-Rebuild simulations and others better matched the results of protocol utilizing the HighTemp-Packer. Both computational protocols result in higher sequence convergence, for wild-type amino acids as well as incorrect amino acid types, than the experimental selection. The two different methods of diversity generation are able to drive escape from the converged energy well for different positions in the interface, indicating that they can each overcome different types of protocol limitations (Figure 17, Figure AI.10-AI.11). For example, positions Ala68 and

Ala70 are converged in the DNA-Rebuild simulations, likely due to the conformation of the

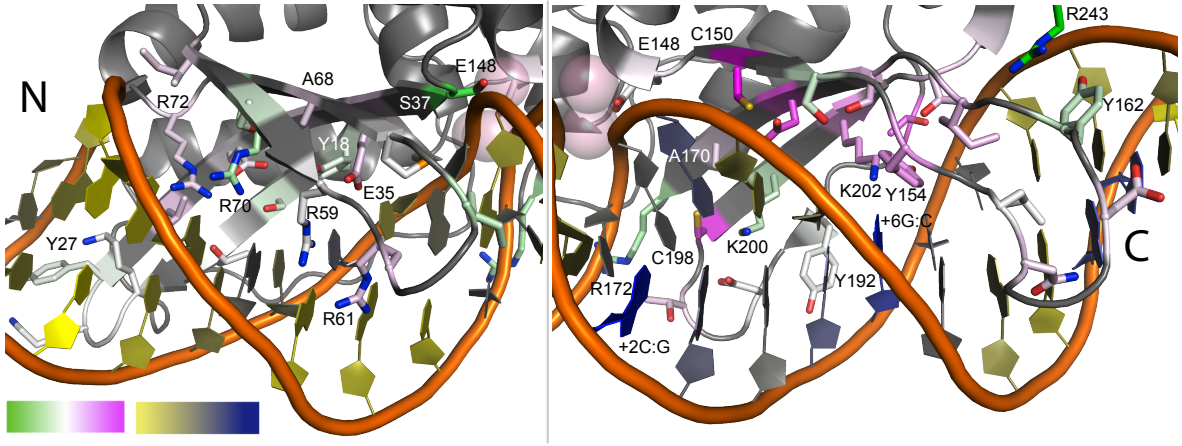


Figure 17. Recovery of experimental data with computational methods. A comparison between the two methods of sequence diversity generation, DNA-Rebuild and HighTemp-Packer is summarized on the structure of I-AniI. The frequency distributions at each of the I-AniI interface positions were compared to the experimental data (Figure 14) by both Euclidean distance and Jensen-Shannon divergence measures (Table 1, Figure AI.10, AI.11). For this illustration, the Jensen-Shannon divergence measure [71] calculated for the DNA-Rebuild method was subtracted from the same calculation completed for the HighTemp-Packer. White is designated as a value of zero, indicating that neither computational method better matched the experimental frequency distribution, green is negative values, indicating that the DNA-Rebuild performed better than the HighTemp-Packer, and pink is positive values, indicating that the HighTemp-Packer performed better than the DNA-Rebuild method. The DNA is colored based on the average RMSD between the DNA-Rebuild simulations and the crystal structure DNA, where yellow is the lowest average RMSD and blue is the highest. The DNA moved farthest away from the crystal structure DNA in the same area that the DNA-Rebuild method performs much less well than the HighTemp-Packer, indicating that the DNA location has a significant effect on the design results.

protein backbone structure. The HighTemp-Packer method was able to generate significant diversity at both these positions that better matched the experimental data. Some positions near the DNA backbone benefited more from the DNA-Rebuild simulation. Positions 37 and 172 show very high convergence in the HighTemp-Packer results, and the experimental data indicates that there should be minimal amino acid preferences here. Both these positions are directly interacting with the DNA backbone in the crystal structure of the complex, and the DNA-Rebuild method was able to reproduce this experimental variation by allowing DNA backbone movement.

The failures of the DNA-Rebuild method are focused in the (+) half of the DNA target site. The interactions with this DNA half-site are implicated in formation of the catalytic complex [61], so it is likely that preservation of the DNA conformation observed in the crystal structure is essential for maintaining activity. Many crystallized protein-DNA complexes contain DNA that is perturbed away from canonical B-form, presumably with a functional purpose. The current implementation of DNA energetics and rebuilding is not yet adequate for capturing the subtleties of these more strained DNA conformations. The DNA-Rebuild method results in low

recovery at several I-AniI positions making (+) half-site interactions that do not show significant variation in the experimental data. For example, position Cys150 is maintained as a cysteine or a serine in the experimental data, and the HighTemp-Packer simulation almost exactly produces the frequencies observed experimentally for these two amino acids. The DNA-Rebuild simulation allows numerous amino acids to be incorporated at this position, as the DNA moves away from the crystal structure conformation. The experimental data for position 150 indicates that maintaining the conformation of the bases in this area is likely critical to catalysis. Additionally, the two most conserved residues in the (+) half-site, Lys202 and Tyr154, are lost in most of the DNA-Rebuild simulations. Figure 17 shows that the DNA is rebuilt in such a way that it moves away from the crystal structure conformation. This non-native DNA conformation allows alternative amino acids to be designed in this area. It is likely that contributions of the DNA conformational state to catalysis in I-AniI are the cause of these inaccurate computational rebuilds. A loss in recovery with the DNA-Rebuild method for other proteins in the benchmark set may similarly be attributable to discrepancies between real and modeled DNA conformational preferences, providing an avenue for improvement of Rosetta's modeling of DNA flexibility.

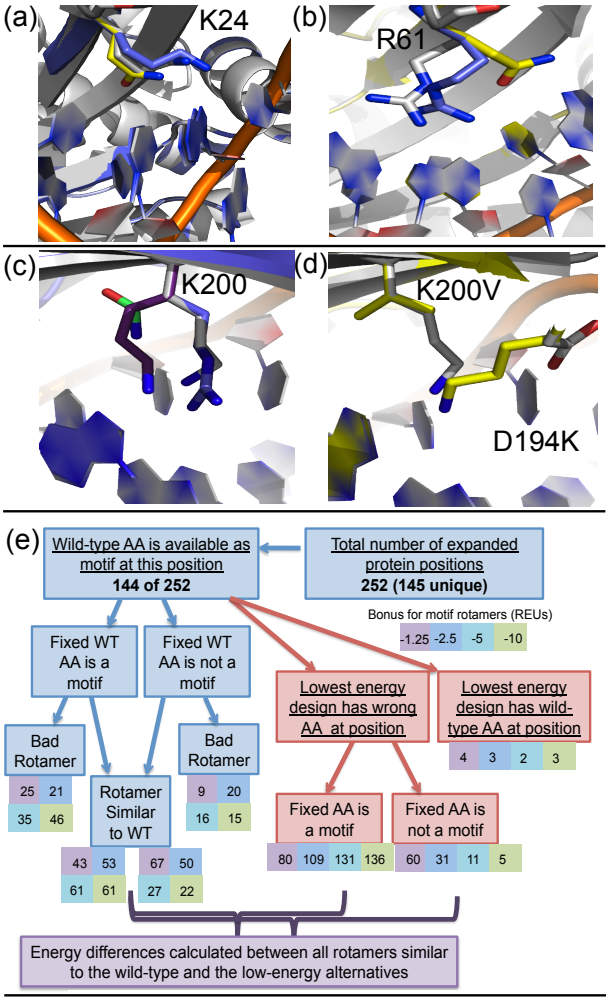
Escaping energetic minima with motif-based sequence constraints

Both of the new protocols for diversity generation fail to recover the experimentally preferred amino acid at some I-AniI positions. One of the essential arginine residues in the N-terminal domain, position 61, is highly conserved as the wild-type amino acid and is not observed as arginine with any protocol. Position 24 is a lysine in the native enzyme, and the enzyme tolerates a lysine or histidine. Neither the DNA-Rebuild nor the HighTemp-Packer recapitulates either of these two possibilities. The previously discussed position 200 is known to be highly active as a lysine (native), asparagine, or arginine, yet none of these amino acids are observed in the computational results.

In order to understand the factors responsible for these mis-designed residues in I-AniI, as well as others in the full sequence recovery set, a modification was made to the previously described protocol for design with motif rotamers. This modified protocol forces amino acid types at each designable protein position to all of the types seen in motifs selected for that

position. For example, if both arginine and lysine motifs passed the search procedure for a particular position the protocol would produce a set of designs with the lysine amino acid type fixed, but not any particular rotamer, at that position, as well as a set with the arginine amino acid type fixed. This sequence constraint can result in sampling of higher energy alternative structures that better match the wild-type protein sequence, and energetic analysis of these forced amino acids has the potential to reveal why those positions are incorrectly designed without the constraint. In addition, this protocol can be used to generate diverse sequences, revealing many potential native-like interactions instead of only the lowest energy one, for seeding experimental libraries

The motif-based sequence constraint method revealed that there is a motif found for every one of the described I-AniI failures. When position 24 is forced to be a lysine, a motif rotamer is incorporated into the design with a very similar conformation to the native lysine (Figure 18a). The competing low-energy Glutamine type is never seen in the experimental interface screen. The difference in total energy between the designs with the lysine and the glutamine is only 0.6 REUs and when compared to all forced motifs, the design with the forced lysine is the second lowest in energy. The dominant energy term disfavoring arginine at position 61 (Figure 18b) is the probability of the amino acid given the backbone conformation (p_aa_pp), having a value of 2.46 REUs for the arginine that is forced with the sequence constraint protocol and -0.92 REUs for the lower energy glutamine type. At position 200, all three of the known, high-activity amino acid types (lysine, arginine, and asparagine) are found to be motifs (Figure 18c). However, none of these types is designed with the standard motif protocols, due to a competing alternative design that incorporates a valine at position 200 and a lysine at the nearby position 194 (Figure 18d).



In order to test this motif-biased sequence constraint protocol on proteins other than I-AniI, it was first necessary to determine which interface positions are likely to be most important for wild-type activity in the absence of experimental data. Given the comprehensive and computationally intensive nature of this protocol it was additionally necessary to limit its use to a subset of designs. The training set was analyzed to determine the residues that are true failures of the design protocol using a set of metrics described in the methods. These mis-designed positions are characterized as failures because they are likely important amino acids, as they are amino acids with significant interaction energy, that are designed to a chemically very different amino acid type. The protocol identified 284 of the 3421 designed protein positions from the training set to be failures, which was further reduced to 252 when additional computational constraints due to protein size were taken into account. These design failures were subjected to the described

Figure 18. Motif-based sequence constraints. (a) Lys24 in the I-AniI interface (native rotamer = white) is mis-designed to a glutamine (yellow). The motif-based sequence constraint protocol revealed that position 24 can be a lysine motif, and the motif residue (blue) very closely matches the native lysine. (b) Arg61 in the I-AniI interface (native rotamer = white) is mis-designed to a glutamine (yellow). The motif-based sequence constraint protocol revealed that position 61 can be an arginine motif (blue). (c) The motif-based sequence constraint protocol showed that position Lys200 in the I-AniI interface (native rotamer = white) can be a motif of any of the three amino acid types previously identified to be active at this position (arginine = blue, lysine = purple, and asparagine = green). (d) The alternative low-energy design that disallows any of the motifs in (c) to be designed at position 200. The native structure is shown in white and the design with K200V and D194K is shown in yellow. (e) Abbreviations: WT = wild-type and AA = amino acid. Flow-chart summarizing the results of the protocol that generates designs with forced amino acid types for each type of motif identified by the motif search. The protocol was completed only for protein positions that were considered to be true failures of the computational methods by a series of analyses. The chart summarizes the motif status, energetics, and rotameric state of the designs at each of these failed positions. Rotamers are considered similar to the wild-type amino acid if they have an RMSD of <0.8. (f) Energy differences calculated between rotamers that resemble the wild-type amino acid that have a motif rotamer incorporated with a bonus and between the incorrectly designed amino acid observed at this same protein position in the lowest energy design, as marked on the flow-chart in (e). The repulsive energy term (fa_rep) stands out at the biggest contributor to the energy difference between these rotamers.

protocol in which the motif residue types are forced at each designable position. This procedure revealed that for 108 of the 252 positions a motif of the same type as the wild-type amino acid is not even available (Figure 18e). For the 144 of these positions where the wild-type amino acid is present in the motifs selected for that position, the number of times that the design actually contains the motif rotamer when the amino acid type is fixed as wild-type was found to range from 68 to 107, depending on the motif scoring bonus. The rotameric state of the amino acid making the motif contact was additionally assessed.

For essentially all of the 144 designed positions where a wild-type motif is available, an alternative design sequence that lacked the wild-type amino acid at that position was found to have a lower energy. These designs with the lowest total energy scores were analyzed to determine the motif status of the mis-designed position. Even for the lowest motif scoring bonus, over half of the positions had a motif rotamer incorporated at the failed position. The components of the energy function were again dissected for each failed protein position by comparing each component from the lowest energy design and from the design with the forced wild-type amino acid, restricting to positions in which the motif rotamer from the forced wild-type simulation was similar to the native rotamer (Figure 18f). The results were significantly different from the previous analyses of this type, as the repulsive score (fa_rep) was found to be responsible for the majority of the energy differences between the forced wild-type amino acid and the alternative low-energy designed rotamer. The rotamer probability term is no longer a major component of these differences. These results suggest that the energy function is favoring sidechains that are less tightly packed, alleviating the clashes recognized in the high repulsive score.

4.4 Suggestions for Improving Modeling of Protein-DNA Interfaces

Human intuition is a valuable tool for assessments of protein interactions [72]. Visual analysis of the designs in the training set was used as an additional metric guiding the process of energy function improvement. A large number of the true failures, as determined by analysis described in earlier sections and the methods, were visually evaluated in order to gain ideas for the necessary next steps in computational method optimization. While there are many reasons a design procedure may result in a non-native amino acid at a protein position, visual analysis of these designs revealed recurrent themes. Four representative design examples are shown in Figure 19a-d. Of these four examples, one is included to demonstrate how not all mis-designs of the wild-type sequence should be considered failures (Figure 19a), one was corrected with the HighTemp-Packer sampling strategy

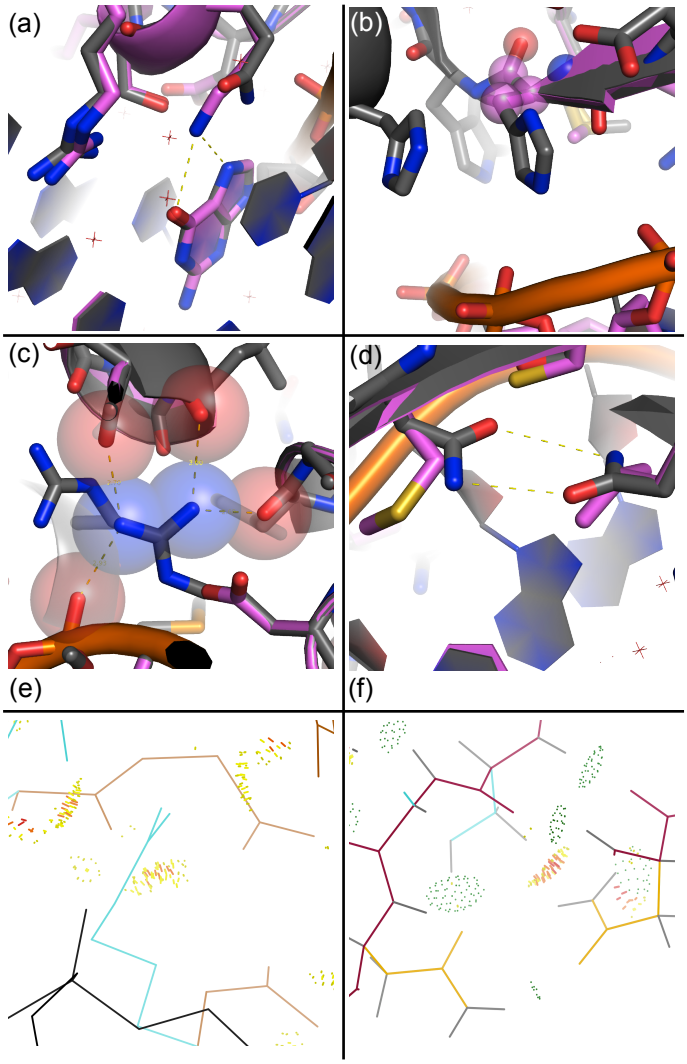


Figure 19. Representative failures of the computational methods. Native structure=gray, designed structure=pink. (a) The designed lysine, making a canonical contact with the guanine nucleotide, is calculated to interact more strongly with the DNA than the wild-type glutamine (Gln39, 1zs4) and no interactions with neighboring protein positions are lost from this substitution. (b) A histidine (His97, 2fl3) is redesigned to an alanine and energetic analysis revealed that the rotamer probability term was mainly responsible for the alanine preference. The HighTemp-Packer method corrects this failure, as the histidine is regained in 71% of the design trajectories, compared to 19% with the DNA-Rebuild method. (c) An arginine residue (Arg432, 1j1v), making multiple contacts to both the protein and DNA backbone atoms, is redesigned to a smaller aspartate residue that makes no favorable interactions. The atoms in the starting crystal structure are very close to each other, and the repulsive clashes cannot be relieved without backbone movement or minimization. (d) A bi-dentate asparagine-asparagine hydrogen bond is lost (Asn70-Asn90, 2ex5). This failure is also due to repulsive clashes with the nearby protein backbone. (e) Amount of atomic overlap Arg432 in the 1j1v crystal structure calculated using MolProbity [74,75,76]. The atomic overlap is shown with yellow and red dots, DNA is black, sidechains are cyan, and the protein backbone is brown. This analysis indicates that the protein backbone and neighboring sidechain residues are clashing with Arg432. Backbone optimization would be required to relieve the clash with the backbone. (f) Atomic overlap (yellow and red dots) between an asparagine residue (yellow) and a hydrogen atom (grey) of the beta carbon of a neighboring serine residue shown in cyan (2ex5, Asn90-Ser68). Hydrogen bonding between this same serine residue and the other asparagine (Asn70-Ser68) of the bi-dentate asparagine pair is shown with green dots.

described in this work (Figure 19b), and the remaining two are the result of the fixed backbone approximation and not optimizing the starting crystal structure in the Rosetta energy function prior to design (Figure 19c, d).

For the three representative cases (Figure 19b-d) where the redesigned sequence is clearly suboptimal to the wild-type sequence, small movements of the backbone of the protein and DNA prior to design would most likely correct the failures. The histidine that was redesigned to an alanine (Figure 19b) was lost because of an excessively high penalty from the rotamer probability term. The energetic contribution of the rotamer probability is dependent on the backbone structure, so subtle movement of the protein backbone would likely correct this failure. For the remaining two cases (Figure 19c, d), the residues being incorrectly designed are all making interactions with the surrounding protein residues. It is possible that these positions provide protein structural stability and thus binding-site pre-organization for these interfaces [73]. The atoms making the primary protein-protein interactions are clashing, as determined by MolProbity [74,75,76], and constrained on multiple sides by the backbone of the protein or DNA, thus prohibiting repacking and instead favoring redesign to relieve repulsion (Figure 19e, f). The findings for these two examples match the results of the motif-based sequence constraint protocol that the repulsive term was the major source of the higher energy of the designs containing the forced wild-type amino acid type (Figure 18f). Optimizing the crystal structures in the Rosetta energy function prior to design is one potential solution to this issue, although this protocol would need to be thoroughly assessed to ensure that it was not generating a bias in the designed sequences for the wild-type amino acids. One way to avoid this artificially generated bias would be to optimize the structures with a different energy function from an external program.

In this work, a number of optimizations to Rosetta have been thoroughly characterized, including energy function improvements and new protocols for sampling diverse design sequences. Limitations of the computation were illuminated, some of which were addressed and others of which still need to be corrected, and a series of methods and analysis tools were developed to increase the ease of such future endeavors. The question of reliability of sequence recovery as a sole metric for energy function improvement was explored in the context of a particularly well-studied enzyme scaffold. Recapitulation of experimental data is a more relevant metric of protein sequence redesign success than sequence recovery, as it removes the biases of

potentially overtraining for recovery of the amino acid states observed in crystal structures and is a more direct measure of the functional effect of allowing a protein sequence to vary. There are many factors contributing to the activity and specificity of DNA-binding or cleaving proteins, such as the transition between the bound and unbound states and the role of neighboring DNA in the formation of the active complex. A crystal structure reveals one state of the interaction complex, and a computational design tool meant to predict sequence changes required to confer certain activities should be assessed with corresponding experimental data, rather than recapitulation of this single, fixed state. Utilizing this combination of experimental and computational benchmarks has revealed several avenues for continuing improvements of the design methodologies. Additionally, the extensive experimental scan completed in this work provides a better understanding of a class of enzymes being actively engineered as gene therapy reagents, and knowledge of the mutability of each position in this particular enzyme will inform future specificity redesign projects.

The Rosetta force field integrates physicochemical energy terms and database-derived potentials in order to guide sampling and selection of low energy amino acid sequences. Similarly, the incorporation of interaction-biased motif rotamers into the standard design process provides a way to integrate the information available in the protein databank with the energetic guidance of the Rosetta force field. The collection of motifs can be considered as a step toward formulating a recognition code [77,78] for protein-DNA interactions. The interactions in protein-DNA interfaces are complex and shaped by the local environment, suggesting that the information contained in motifs is best utilized in combination with a tool for assessing the likelihood of a given motif in the context of the entire interaction complex. The method described in this work builds on a previous approach in which the motif interaction is held constant as the protein backbone is remodeled to stabilize the desired contact [47,64]. Camacho and coworkers have recently described an alternative computational method for investigating this recognition code that combines homology modeling and molecular dynamics simulations to predict changes in binding affinity for zinc-finger mutants [79]. One significant advantage of this approach over the current Rosetta methods is that explicit waters were simulated at the interface, allowing for improved modeling of water-mediated interface contacts. The incorporation of explicit water into the Rosetta protein-DNA interface design calculations is currently under study.

While the addition of the motif rotamers improved the results of the Rosetta design protocol, the optimization of the force field resulted in an even more significant improvement. Indeed, as the force field was iteratively improved, the optimal value for the motif bonus term decreased, suggesting that the new and modified energy terms were able to preferentially reward native-like protein-DNA interactions. While encouraging, these improvements – when applied in the context of the standard fixed-backbone design simulation – did not enable successful recapitulation of the variability seen in our I-AniI experimental dataset. To explore the potential role of DNA backbone flexibility, we integrated a recently described method [53] for generating diverse DNA conformations into our design protocols. Most other programs for protein-DNA interface design, such as FoldX [80], use a fixed backbone model of the DNA. While preliminary DNA minimization was available in older versions of Rosetta [54], this new implementation of DNA flexibility is significantly more flexible and provides for greater DNA backbone movement (due to the fact that Monte Carlo fragment rebuilding simulations sample a much larger conformational space than gradient-based minimization initiated at crystal structure conformations). Both this new method of sequence diversity generation and the HighTemp-Packer method, defined by an increase in the final temperature used by the simulated annealing algorithm, improve recapitulation of the experimental dataset over standard Rosetta methods (Figure 17).

In contrast to protein sequences generated by computational design, the primary function of the amino acids in a protein-DNA interface is not always the stabilization of the lowest energy state or the tightest possible binding. There also may be a range of binding affinities tolerated for maintaining interface functionality. The wild-type amino acid sequence may not always be the most energetically optimal sequence position at the designed position (Figure 19a). It is challenging to determine whether the seemingly native-like interactions in the design are really compatible with the activity of the protein-DNA complex. Native complexes are evolved for many functions other than tight binding. The only way to fully assess the viability of the mis-designed amino acids is through experimental characterization. There are several positions in the I-AniI interface where the wild-type amino acid is not the most optimal (Figure 14). For example, position 18 has a significant preference for tryptophan over the wild-type tyrosine and the previously discussed position 200 shows high experimental recovery of arginine instead of the wild-type lysine. In these two cases, the preferred amino acid likely confers an increased

selective advantage through tighter substrate binding or catalytic complex formation. While these positions are somewhat tolerant of substitutions, they differ from the many highly tolerant positions in the I-AniI interface in that they display a significant preference for a particular amino acid type, rather than allowing all amino acid types equally. A successful computational design tool would capture these non-native energetic preferences, while predicting a lack of preference at the most flexible positions. While it is currently challenging to determine which classes of interface mutations are systematically mis-predicted due to the limited size of our experimental dataset, we expect that recent work combining next-generation sequencing technology with protein selection [81] will revolutionize studies of this sort that attempt to correlate protein mutations with functional characteristics.

The goal of our work is to develop protocols with clear utility for future design projects. Minimizing the starting structure into the native energy well to alleviate predicted clashes in starting structures (Figure 19) is likely to artificially enhance sequence recovery by biasing toward the wild-type state. Without proper benchmarks, preferably experimental data, it would be challenging to ensure that this over-optimization of the native state was not biasing the results. In light of the experimental data collected for I-AniI that revealed that a number of interface positions tolerated multiple amino acid types, it is likely that the relatively high sequence recovery of 50% is due to an over-optimization for the native sequence in the context of the rigid, fixed-backbone sequence design simulations. While native sequence recovery has proven to be a powerful metric for optimization of protein design scoring functions, its use as the sole benchmark for protein design sampling algorithms would likely penalize the greater exploration of backbone diversity necessary for successful design toward novel DNA target sites. The experimental data is even an underestimate of the acceptable sequence diversity, since only one position is being allowed to change at a time. Varying multiple positions simultaneously would likely show even less conservation of the wild-type sequence due to correlated changes. Computational protocols producing 100% recovery of the wild-type sequences would almost certainly be useless for design purposes. Instead it would be best to perfectly recover the amino acids forming essential interactions in the protein-DNA interface and have low recovery and multiple solutions generated for the more malleable positions.

Developing a way to perturb the starting crystal structure on both the protein and DNA side, without biasing toward the native energy minima, will be important for correcting the

failures identified from the sequence recovery benchmarks (Figure 19). There are a number of possible methods to potentially adapt to provide an alternative method of DNA movement that is less extreme than the fragment insertion protocol tested here [65,82]. Both the loss in recovery when using the DNA-Rebuild method as well as the comparisons to experimental data indicate that less conformational freedom of the DNA is likely to produce higher sequence recovery. However, DNA movement is essential for design of new DNA sequences and for predictions of energetics and specificity involving indirect readout [83,84], so it is important to develop a reliable method for accomplishing this goal. Adding protein backbone flexibility will also be necessary for improving recapitulation of experimental data and generating diverse designed sequences [85,86]. Flexible loop regions of protein-DNA interfaces could benefit from combining the motif-based approach described here with the previously published method that rebuilds protein backbones to accommodate rotamers that can form motif interactions [47]. The results of the simulations completed with the HighTemp-Packer showed promising recapitulation of the variation observed in experimental data. However, the loss of some of the strong motif-like interactions of I-AniI when using this approach suggests that incorporation of the motif information could further enhance the method. One potential way to increase the ease of utilizing the motif information, especially for systems other than protein-DNA interfaces, is to incorporate the data about distances and angles of interactions into a knowledge-based contact potential scoring function [87]. For current design applications we suggest an approach that combines subtler DNA backbone optimization with the HighTemp-Packer and motif rotamers. We hope that these proposed improvements, in conjunction with the newly developed methodologies and analysis tools, will accelerate the progress of future design projects.

Chapter 5

Mining Endonuclease Cleavage Determinants in Genomic Sequence Data

Homing endonucleases have great potential as tools for targeted gene therapy and gene correction, but identifying variants of these enzymes capable of cleaving specific DNA targets of interest is necessary before the widespread use of such technologies is possible. We identified homologues of the LAGLIDADG homing endonuclease I-AniI and their putative target insertion sites by BLAST searches followed by examination of the sequences of the flanking genomic regions. Amino acid substitutions in these homologues that were located close to the target site DNA, and thus potentially conferring differences in target specificity, were grafted onto the I-AniI scaffold. Many of these grafts exhibited novel and unexpected specificities. These findings show that the information present in genomic data can be exploited for endonuclease specificity redesign.

5.1 The Evolution of LAGLIDADG Homing Endonucleases

Homing endonucleases (HEs) are DNA cleaving enzymes encoded by an open reading frame (ORF) located within an intron or intein. These enzymes are highly specific for target sites located on homologous alleles that lack the endonuclease ORF and intervening sequence. The homing endonuclease protein introduces a double-stranded break at this target site that then stimulates the homologous recombination DNA-repair pathway. Repair via recombination leads to duplication of the selfish mobile element containing the HE ORF [22,88]. Harnessing the natural potential of these enzymes through the reprogramming of their substrate specificity [1,58,62] will help drive forward the rapidly expanding areas of genome engineering and gene therapy [11].

Recent work on adapting homing endonucleases for use in such biotechnology applications has focused on the LAGLIDADG family, whose genes are primarily found within archaea, algal chloroplasts, and the mitochondria of fungi [12]. These proteins are so named for the conserved sequence of amino acids in the alpha-helices that separate the N and C terminal

halves of the pseudo-symmetric monomeric LAGLIDADG homing endonucleases (LHEs) and form the binding interface of homodimeric LHEs. These conserved helices include essential catalytic acidic residues that catalyze the cleavage of the target DNA between two scissile phosphates, separated by four base-pairs referred to as the central four [22,89]. Crystal structures of both homodimeric and monomeric members of this protein family reveal a well-conserved, canonical fold as well as a similar curvature of their approximately 20-22 base-pair long DNA substrate [90,91,29]. Specific interactions with this target DNA are made by residues in two β-sheets that flank the central four bases and traverse the major groove of both the plus and minus DNA half-sites [88]. Homing endonuclease genes (HEGs) are selfish genetic elements, and their continued existence relies on their ability to sustain homing in the face of genetic drift and to invade new, albeit related, host organisms. The extended β-sheet topology for protein-DNA interactions provides this flexibility by allowing for low fidelity at some positions in the interface, most often at the wobble positions in the host gene sequence, while maintaining an overall high level of specificity due to the length of the DNA substrate [24,92].

The pair of structurally characterized HE homologues I-CreI and I-MsoI provides a prime example of the extensive flexibility of this protein scaffold. These two enzymes cleave nearly identical target sites, yet they share only 38% sequence identity with only 5 of the 25 DNA contacting residues conserved [90]. Such divergent evolution can facilitate the acquisition of alternative biochemical functionalities, often with a benefit to the host organism. For example, the endonuclease I-AniI can act as a maturase, assisting in the splicing of the intron containing its own coding HEG [93]. As the selective pressure following HEG invasion is not high, genetic drift can also result in the loss of endonuclease function; the I-AniI homologue BI3 maturase acts exclusively as a maturase, having lost all endonuclease activity due to mutation of one of its catalytic glutamates to a lysine. While the BI3 maturase no longer functions as an endonuclease, and only shares approximately 50% sequence identity with I-AniI, it was observed to still bind the I-AniI target site DNA with high affinity [94].

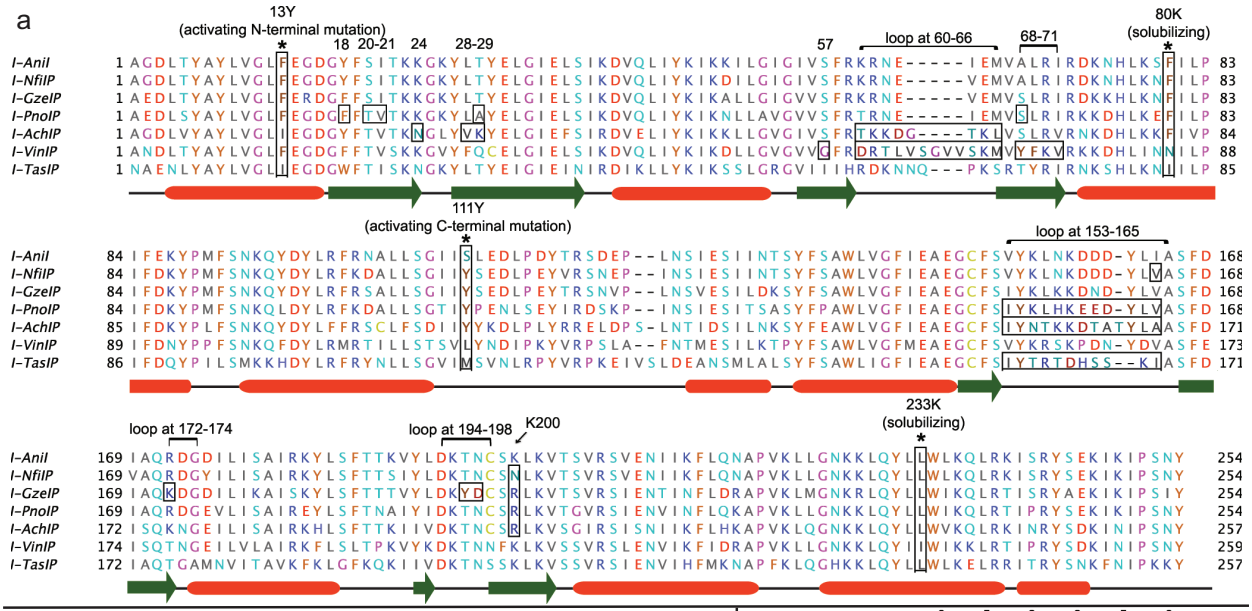
In this era of genome-wide sequencing new LAGLIDADG ORFs are being identified rapidly [21,95], including a large number of homologues of enzymes that are already being engineered for gene targeting applications (such as I-AniI). For most protein families, determining the native substrate and specificity of newly identified proteins is usually quite labor intensive. However, target site identification for homing endonucleases can be achieved with

careful analysis of the sequence flanking the mobile element containing the HEG. As a result, we can compare amino acid substitutions and putative target sites between various homologues. Analysis of endonuclease sequence alignments revealed a number of amino acid differences between enzymes that were predicted to cleave similar target sites, leading us to question what effect these mutations have on endonuclease activity and specificity. The work described here addresses this question for one LAGLIDADG endonuclease, I-AniI.

I-AniI is one of the most thoroughly characterized homing endonucleases, and its particular biochemical properties make it a promising candidate for gene therapy applications. Importantly, degradation of the enzyme's activity from genetic drift has been artificially reversed using directed evolution methods, resulting in an enzyme with high activity in human cell lines, an essential feature for a gene therapy reagent [30]. Additionally, the specificity of the enzyme, and the effect of base-pair substitutions on kinetic activity, has been acquired for every position of the target site, and variants with novel target site specificities have been generated by computational redesign of the protein-DNA interface [31,61]. In this study we dissect the differences in specificity between close homologues of I-AniI.

Identification of I-AniI homologues and their putative target sites.

BLAST searches against the NCBI non-redundant (nr) database using the I-AniI sequence identified a group of homologues. The ORFs encoding homologues with greater than 47% identity to I-AniI (Figure AI.13) were identified from a variety of fungal species, at insertion sites within an intron of the same host gene as I-AniI (cytochrome B, or 'COB'). Since residence in that particular intron suggests that at one point these enzymes were able to cleave the same or a closely related intronless allele, it is likely that they have activity on a target site similar to that of I-AniI. We found that in alignments with less than 40% identity it was significantly more challenging to identify the flanking target sites, as these enzymes were inserted in a different location in the mitochondrial genome and the exact target site boundaries were unclear. We therefore focused our analysis on homologues that were unambiguously determined to reside in the same intron as the I-AniI ORF. A multiple sequence alignment of all such I-AniI homologues available at the time of publication is shown in Figure AI.13 and the subset of these enzymes that were analyzed in this study are shown in Figure 20a.



Homologue Name	Percent Identity to I-AniI	Percent Similarity to I-AniI
I-NfiIP	93%	98%
I-GzeIP	84%	93%
I-PnoIP	78%	91%
I-AchIP	71%	86%
I-VinIP	64%	79%
I-TasIP	57%	74%

Figure 20. I-AniI homologues and predicted cleavage sites. a) Multiple sequence alignment of ORFs encoding the putative (indicated by suffix P) LAGLIDADG homing endonuclease homologues of I-AniI that were experimentally explored in this study. Groups of amino acids that were transferred to the I-AniI scaffold are boxed with a black border and labeled above, as are the positions of activating and solubilizing mutations (both additionally denoted by an asterisk). Secondary structure elements are identified from the 2QOJ crystal structure definitions, with α-helices indicated by red boxes and β-sheets by green arrows. See Table S2 for a complete list of the mutations made for every hybrid protein. **b)** Percent identity and similarity of each homologue to I-AniI. **c)** Putative target sites identified for each homologue by examining the area surrounding the homologue ORF and comparing to the I-AniI target site. Positions in endonuclease target sites are identified in relation to the predicted center of the site, with positions on the left, or (-) half-site, designated as -10 to -1 and positions on the right, or (+) half-site, designated as +1 to +10. Wobble positions in the COB gene are boxed in black and labeled.

Examination of the DNA flanking this intron in homologues reveals nucleotide sequences that closely match that of the I-AniI substrate (Figure 20c, Figure AI.14, showing the relationship between the intronic sequence, target site, and endonuclease ORF). However, aligning these putative sites with that of the wild-type reveals some variation, mainly at the wobble positions in the codons of the COB protein. There are a number of positions in these homologues at which the amino acid identity differs from that of I-AniI. Some of these residues have the potential to directly interact with the substituted bases in these putative target sites while others are non-conservative and do not neighbor any predicted base changes. In order to understand the effect of these mutations on the specificity and activity of I-AniI, we constructed

hybrid endonucleases by transferring amino acids observed in the homologue alignments to I-AniI.

5.2 The Activity and Specificity of Hybrid I-AniI Endonucleases

Enzyme hybrids based on each I-AniI homologue were made by transferring specific amino acids to the well-characterized I-AniI scaffold, and the local specificity shifts resulting from these mutations were analyzed with *in vitro* DNA cleavage assays.

Generation of hybrid I-AniI homologue endonucleases

The homologues selected for testing (Figure 20b) have sequence identities ranging from 57% to 93% relative to I-AniI (Figure 20a, Table AI.5) and are predicted to cut very similar sites (Figure 20c). Thus, a large proportion of their protein mutations are likely to be neutral and the result of genetic drift rather than arising from selective pressure to act on different substrates. We hypothesized that any specificity differences between members of the I-AniI subfamily might correspond to changes in individual DNA binding residues that have arisen during their evolutionary divergence, and that these changes might illustrate new strategies to alter the recognition specificity of LHEs. To test this concept, we chose to transfer subsets of amino acids from the homologue sequences to I-AniI and to experimentally test the activity and specificity of the resulting hybrids.

The expression of four full-length homologues was attempted in order to compare their specificity and activity to that of the hybrid endonucleases. Of these four homologues, only I-NfiIP was soluble and its specificity profile is available in Figure AI.15. Specificity profiles for the other three endonucleases (I-AchIP, I-VinIP, and I-TasIP) could not be obtained due to poor expression, with all expressed proteins observed to be insoluble by SDS-PAGE. This challenge of reliably characterizing many diverse homologues can potentially be alleviated by hybrid generation methods, where mutations predicted to affect properties of interest are transferred onto a stable, well-studied protein scaffold.

To identify possible mutations for transfer from homologues to I-AniI, the amino acid sequence of each homologue was threaded onto the I-AniI crystal structure, and mutations were

grouped into four categories - the protein surface distant from the bound DNA, direct interface contacts between DNA and protein, the protein core, and the peptide linker between protein domains – depending on their location. The most interesting of these groups, from the perspective of specificity and activity-altering potential, are those that are located in the protein-DNA interface and those in the enzyme core that are immediate neighbors to interface changes. We selected a subset of these substitutions (Figure 20, Table 2) for transfer to the I-AniI scaffold. The groups of amino acids chosen were either adjacent to a change in the putative target site of the homologue or were non-conservative substitutions directly in the protein-DNA interface that we hypothesized could alter the interaction of the enzyme with the DNA. The following sections describe the activities and specificities of the hybrid endonucleases that were characterized, with the results summarized in Table 2.

Table 2. Summary of altered specificities and activities for I-AniI hybrids. Variants are grouped into categories (C-terminal loops, K200, Central 4 loops, Core Mutations) dependent on the location and theorized role of the mutations transferred to the I-AniI scaffold. The base activity column indicates whether the variant was made with either of the activating F13Y or S111Y mutations. Quantitative activities, cleavage plots, and additional information on each variant are available in Table AI.6 and Figures AI.16-AI.20.

Variant	Mutations from I-AniI	Base Activity	Tested Positions	Effect on specificity and activity
C-terminal Loops				
I-PnoIP	V153I, N157H, D159E, D160E, I164V	F13Y	+8, +9, +10	▪ Specificity pattern similar to I-AniI ▪ Activity increased for most substitutions
I-AchIP	V153I, K155N, L156T, N157K, D160T, inserted A after 160, D161T, I164A	F13Y	+7, +8, +9, +10	▪ Novel +7A specificity ▪ Loss of specificity at +8, +9, and +10
I-TasIP	K155T, L156R, N157T, K158 deletion, D160H, D161S, Y162S, L163K	F13Y	+7, +8, +9, +10	▪ Shifts favoring +7A, +8G, and +9C
K200				
I-NfiIP	I164V, K200N	F13Y	+3, +4	▪ Novel +3C specificity ▪ Similar specificity to I-AniI at other substitutions
I-PnoIP	K200R	F13Y	+3, +4, +5	▪ Increased cleavage of +3G and +4T
Central 4 Loops				
I-GzeIP	R172K, T196Y, N197D	F13Y	+2, +3	▪ Specificity pattern similar to I-AniI ▪ Activity increased for most substitutions
I-AchIP	K60T, R61K, N62K, E63D, inserted G after 63, I64T, E65K, M66L	S111Y	-2	▪ Purines preferred over pyrimidines
I-VinIP Loop	S57G, K60D, N62T, E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	-2	▪ -2C favored over WT -2T ▪ Matches -2C target site prediction
Core Mutations				
I-VinIP w/o core	A68Y, R70K	WT	-6, -5	▪ Minimal activity observed on -6 and -5 targets for variant with interface mutations only ▪ Activity recovered with addition of core mutations L69F and I71V
I-VinIP w/ core	A68Y, L69F, R70K, I71V	WT		
I-VinIP-S111Y w/o core	A68Y, R70K	S111Y	-6, -5	▪ Novel -6T specificity ▪ Enhanced cleavage activity due to core mutations
I-VinIP-S111Y w/ core	A68Y, L69F, R70K, I71V	S111Y		
I-VinIP Loop	S57G, K60D, N62T, E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	-6, -5	▪ Activity increased further relative to I-VinIP-S111Y w/ core
K24N/L28V/T29K	K24N, L28V, T29K	WT	-8	▪ L28V core mutation enhances activity ▪ Computationally predicted K24N and T29K highly specific for -8G ▪ Matches -8G target site prediction
K24N/T29K	K24N, T29K, lacking L233K	WT		
I-PnoIP N-terminal transfer*	Y18F, S20T, I21V, T29A, A68S	S111Y	-6, -5	▪ -5T substitution preferred over WT -5A ▪ Matches -5T target site prediction

*Quantitative data and cleavage plots for this variant, included as an additional example of transferred mutations conferring specificity towards the putative homologue target site, are given in the supplement.

Loop transfers at the edge of the DNA target site

A surface-exposed, C-terminal loop near one end of the I-AniI protein scaffold contacts the final four base-pairs (corresponding to positions +7 to +10, in the (+) half of the target site (Figure 21e)), and extends, approximately, from residue 153 to 165. I-AniI displays relatively low specificity in this region compared to the rest of the target site, showing little preference for one nucleotide over another, especially at positions +8 to +10 (Figure 21a). This reduced specificity is presumably due both to the heightened flexibility of the loop (which displays relatively few packing interactions against the underlying core of the protein) and also to the lower number of direct protein-DNA interactions. However, subtle effects on activity are still observed for different substitutions at these positions. Furthermore, it has been previously shown that specificity can be significantly increased at the +8 position with structure-based computational redesign of the enzyme, indicating that it is possible to generate enzymes with alternative and higher specificities in this area.

This loop is one of the more variable regions of the protein-DNA interface amongst the selected homologues. Three different loops were transferred to the I-AniI scaffold and tested for their effect on the enzyme's ability to discriminate between target sites with single base pair substitutions in the +7 to +10 range (Figure 21a, Figure AI.16). These loops, derived from the I-PnoIP, I-AchIP, and I-TasIP endonucleases, have varying levels of similarity to the I-AniI loop and all three are different lengths, as is highlighted in the sequence alignment of this region in Figure 1a. The hybrid proteins were made in the context of the F13Y mutation in the I-AniI scaffold (a sequence change that increased catalytic activity, discovered during directed evolution experiments on that protein [30]. The relative activities of the enzyme hybrids using various substrates were therefore compared to I-AniI-F13Y activity on the same substrates. I-PnoIP has the highest sequence identity to I-AniI of the three enzymes, both overall and in the transferred C-terminal loop. This loop is the same length as the analogous I-AniI loop, but has changes in five of the twelve amino acids. V153I is a conservative substitution that points into the core of the protein, D159E and D160E do not appear to interact with the substrate, and N157H and I164V are located within contact distance to the bound DNA and thus are most likely to alter the specificity of the interface (Figure 21e). We tested activity of the I-PnoIP hybrid (in

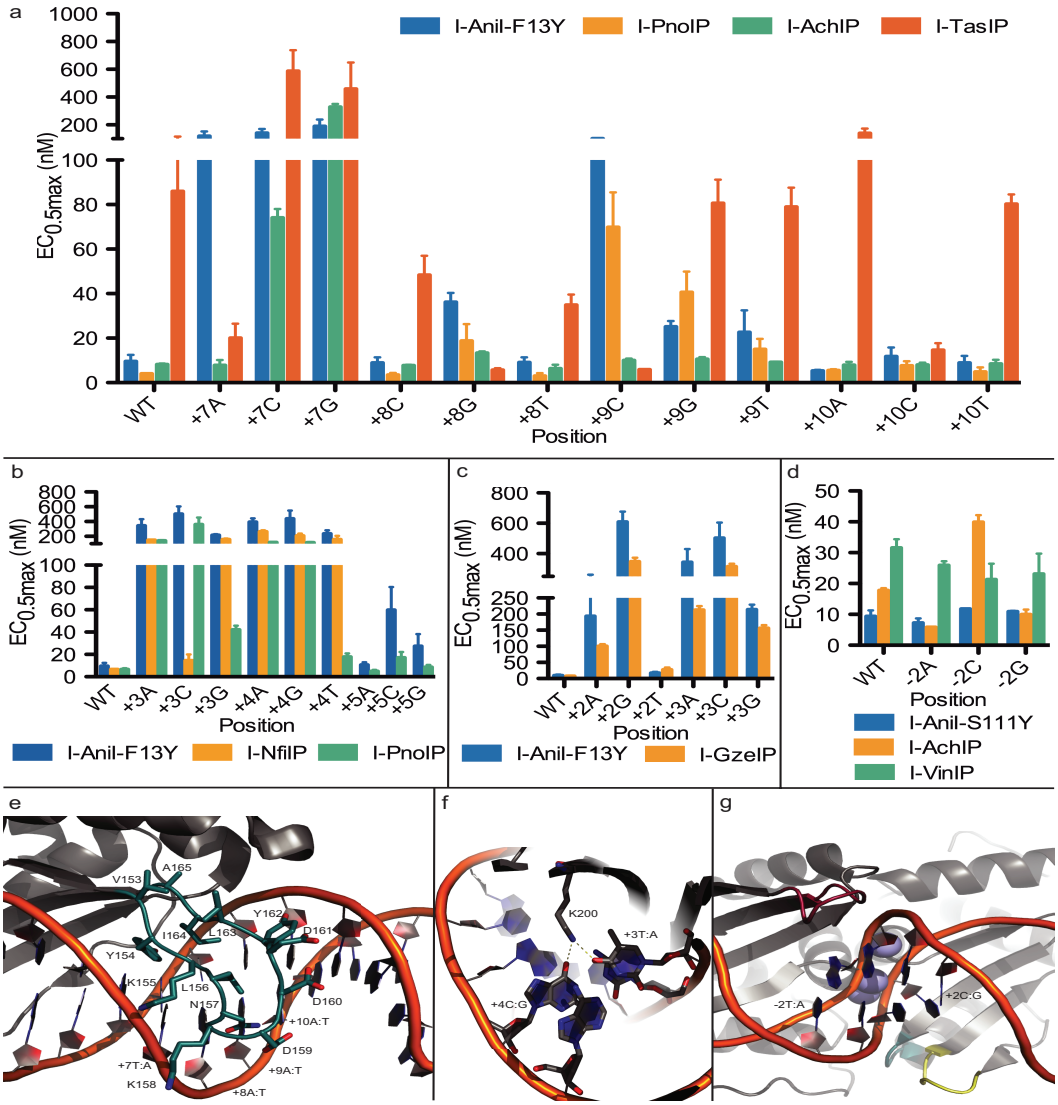


Figure 21. Transfer of loops and interface residues results in new specificities. Error bars in all panels are standard errors from the mean (SEM). See Table AI.6 for a complete list of the mutations made for every variant protein. **a)** Cleavage efficiencies as quantified by $EC_{0.5max}$ (nM) at the +7 to +10 positions contacted by the C-terminal distal loop. Loops transferred from the I-AchIP and I-TasIP homologues are shown to significantly alter the specificity profile of I-AniI-F13Y, most notably introducing new activity for +7A. **b)** Cleavage efficiencies at +3, +4, and +5 for the K200 variants tested. In comparison to I-AniI-F13Y, transferring the K200R found in I-PnoIP markedly improves +3G and +4T cleavage, while K200N from I-NfiIP improves +3C cleavage. **c)** Results from homologue I-GzeIP-derived transfers of the two C-terminal loops near the central four nucleotides show increased activity at positions +2 and +3. **d)** N-terminal loop transfers from I-AchIP and I-VinIP demonstrate altered specificities at the -2 position. The I-VinIP loop was tested in the context of additional mutations from the I-VinIP homologue that were incorporated for a related experiment to analyze the effects of core mutations on position -6 specificity and activity (Figure 3a, panel 5). The reduced activity on all target site substitutions at the -2 position is presumably due to a lower activity on the wild-type site arising from these other amino acid changes. **e)** The C-terminal distal loop between V153 and A165 as seen in the solved I-AniI crystal structure (2QOJ [29]) is colored in blue. **f)** Residue K200 forms direct hydrogen bonds with wild-type target site bases +3T and +4C. **g)** Three loops in I-AniI that contact the central four target site positions are shown. The loop in the N-terminus roughly spans residues 60 to 66 (red) and contacts the -2 position. Two C-terminal loops potentially affect positions +2 and +3: one from 172 to 174 (teal) a second from 194 to 198 (yellow). Significant variation in both the sequence and length of these loops is seen among I-AniI homologues.

which the sequence of the I-AniI loop was changed to that of I-PnoIP) on the singly-substituted target sites from +8 to +10 and found to have a very similar pattern of specificity to I-AniI-F13Y, but shows an overall slight increase in activity on the majority of tested substrates (Figure 21a). I-AniI-F13Y cleaves the +8T substitution identified in the putative I-PnoIP target site (Figure 20c) and the few changes made to the loop of this hybrid enzyme did not improve specificity for the thymine nucleotide.

The loop transferred from I-AchIP is one amino acid longer than the I-AniI loop and only four of the thirteen amino acids are conserved between the two regions. This hybrid showed activity equivalent to that of I-AniI-F13Y (Figure 21a), however its specificity profile is significantly altered. Positions +8 to +10 no longer display even the low levels of specificity observed for I-AniI and position +7 now allows an adenine nucleotide in addition to the wild-type thymine. This new specificity observed at position +7 was not previously accomplished using the I-AniI protein during either computational redesign or directed evolution experiments. Therefore, the information derived from the identification and exploitation of homologous endonucleases can be used to create enzyme variants with additional functional capabilities.

The loop derived from the I-TasIP homologue has only one residue in common (Y154) with that of I-AniI and is one residue shorter. The hybrid protein shows remarkably high specificity that differs significantly from that of the wild-type I-AniI. Like the I-AchIP loop hybrid, this I-TasIP transfer also displays activity against the substrate with an adenine at position +7 that is comparable to the wild-type enzyme with the wild-type +7T substrate (Figure 21a). However, unlike the I-AchIP loop hybrid, it displays significantly reduced activity on the wild-type thymine at +7, as well as on the other two +7 singly-substituted sites. In addition, with the inclusion of the I-TasIP amino acid changes to the loop, the +8G and +9C substitutions shift from being the least favored nucleotides (with the wild-type loop) to the most favored. Both the I-TasIP and the untested I-VinIP homologues contain an L156R mutation in their C-terminal loop. This mutation is likely one source for the preference of a guanine at the +8 position, found in the putative target site of both homologues. Due to the difficulty of modeling protein-DNA interactions in a flexible loop, identification of such loop changes that confer novel specificities is key to designing endonucleases.

Sequence changes in the middle of the protein-DNA half-site

Surface-exposed loops are generally considered to be more flexible than secondary structure elements, such as β-sheets. The high degree of variation between homologues in the C-terminal loop described in the previous section, as well as the low level of substrate specificity observed under I-AniI loops [31,61], attest to this flexibility and tolerance to mutation in both partners of the protein-DNA interface. In contrast, the identities of protein residues located in the relatively inflexible β-sheet region of a LAGLIDADG interface are predicted to be more influential on target site activity and specificity.

Residue lysine 200 of I-AniI is located in the center of the β-sheet in the C-terminal domain of I-AniI. This lysine forms direct hydrogen bonds with positions +3 and +4 of the target site (Figure 21f) and likely contributes to the high specificity of I-AniI at both positions. As indicated in Figure 20a, four of the eight homologues studied have a mutation to arginine at position 200, and one, I-NfiIP, has a mutation to asparagine. I-NfiIP has only one other mutation in the C-terminal interface, and the effect of a K200N mutation in I-AniI was tested in the context of this I164V mutation. This I164V mutation was also present in the I-PnoIP C-terminal loop, where it was found to have little effect on activity in the context of the other loop mutations (Figure 21a).

I-AniI hybrids with either mutation to lysine 200 maintain some activity on the wild-type substrate, albeit to different degrees (Figure 21b). The arginine substitution, seen in I-PnoIP and three other homologues, yields slightly more activity on all tested target sites than the corresponding I-AniI-F13Y enzyme, however it also results in an overall reduced specificity; in particular the cleavage of the +3G and +4T targets is significantly enhanced (Figure 21b, Figure AI.17). This observed relaxation of specificity might result from the increased length of the arginine sidechain. The K200N substitution derived from I-NfiIP increased the activity on the +3C site by more than 30-fold, to levels close to that of the wild-type enzyme on the wild-type +3T site (Figure 21b). For the remaining sites tested, the I-NfiIP K200N variant had a similar specificity pattern to I-AniI-F13Y. The comparatively low levels of cleavage activity on the purine bases at position +3, independent of the identity of the residue tested at position 200, may reflect the requirements of sequence-dependent DNA bending on catalysis.

Loop transfers at the center of the DNA target site

The central four bases of many LAGLIDADG endonuclease target sites are distorted away from B-form DNA, due to a greater degree of bending at these nucleotide positions. This region of the DNA is not in direct contact with the enzyme, yet it displays significant sequence preferences that are presumably the result of indirect readout of DNA conformational requirements. While certain DNA sequences within this region of the target may be completely disallowed because of their inability to deform to the necessary conformation, the protein sequence surrounding this region should have some influence on the specificity of the base substitutions that are conformationally accessible. Identifying areas in the protein that can influence the specificity of the central four base-pairs is valuable not only to increase the pool of potential starting targets for further engineering, but also because increasing our understanding of how to control specificity in this region is important for more general redesign of homing endonuclease cleavage specificities.

Loops on both domains of the endonuclease contact the central four positions: one in the N-terminus that extends from approximately position 60 to 66 in I-AniI, one in the C-terminus from positions 172 to 174 that interacts with the phosphate backbone of the central four, and a second C-terminal loop that spans approximately positions 194 to 198 (Figure 21g). There is significant variation in the sequence, as well as the length, of these three loops between homologous enzymes (Figure 20a). Three loop transfer variants were produced: one incorporating sequence changes from both of the C-terminal loops together, made in the F13Y background, and one from each of the two homologues with N-terminal loop changes, both made in the S111Y background. The effects of the C-terminal transfer were tested on positions +2 and +3, while the N-terminal transfers were analyzed on -2.

I-GzeIP was chosen for transfer of C-terminal loop mutations because, while its putative target site is identical to I-AniI, the amino acid changes in this homologue are non-conservative and structurally proximal to the central four base pairs. Amino acid mutations made to the two C-terminal loops – R172K, T196Y, and N197D – resulted in an enzyme with slightly increased activity on the wild-type substrate, with the wild-type cytosine at position +2 still preferred over other substitutions (Figure 21c, Figure AI.18). This hybrid enzyme showed a modest reduction in activity for a thymine substitution at this position, the next most-favored nucleotide for the wild-

type enzyme. Both purine substitutions remain unfavorable for this mutant, however the I-GzeIP hybrid shows higher activity against these sites compared o the wild-type enzyme. The I-GzeIP enzyme was also tested on the +3 singly-substituted targets, and although it maintains the same order of substrate reactivity, it has an overall increased activity on all three alternative substrates.

The I-AniI-S111Y wild-type enzyme has low specificity at position -2, and homologue N-terminal loops neighboring this position were transferred to test for sequence effects on specificity in this region. Both N-terminal transferred loops are longer than the wild-type I-AniI loop; the I-AchIP derived loop is extended by one residue and the I-VinIP loop by five. The I-AchIP loop displays a preference for purines over pyrimidines at the -2 position, cleaving the wild-type thymine with slightly reduced activity than the wild-type enzyme and cleaving the cytosine substitution significantly more poorly than any other nucleotide (Figure 21d, Figure AI.18). Activity on the guanine and adenine counterparts was maintained across the hybrid and wild-type enzymes. Of all analyzed homologues, the I-VinIP ORF is the only one with substitutions in the central four positions of the predicted target site (Figure 20c), with a switch of thymine to cytosine at position -2. The hybrid protein incorporating the I-VinIP loop was found to favor a cytosine nucleotide at position -2, thus corroborating the target site prediction.

Mutations to residues in the protein core

While interface mutations are most likely responsible for specificity and activity shifts, core mutations proximal to these could also contribute to these properties. In our final experiments, we tested whether core mutations could have a substantial effect on enzyme activity and lead to shifts in specificity that are not achieved with interface mutations alone (Figure 22).

Two I-AniI protein-DNA interface mutations, A68Y and R70K, were transferred from I-VinIP. R70 interacts with a guanine nucleotide at position -6 (Figure 22c). The activity of this enzyme on variants at positions -6 and -5 is increased by the addition of two core mutations, L69F and I71V. While the inclusion of these core mutations in hybrids both with and without the S111Y activating substitution led to activity enhancement, the effect was more dramatic with the latter variant (Figure 22a). The hybrid containing both interface and core changes has relaxed specificity at position -6 compared to I-AniI-S111Y, now cleaving the previously inaccessible -6T in addition to the wild-type -6G. An additional hybrid was made by transferring a loop

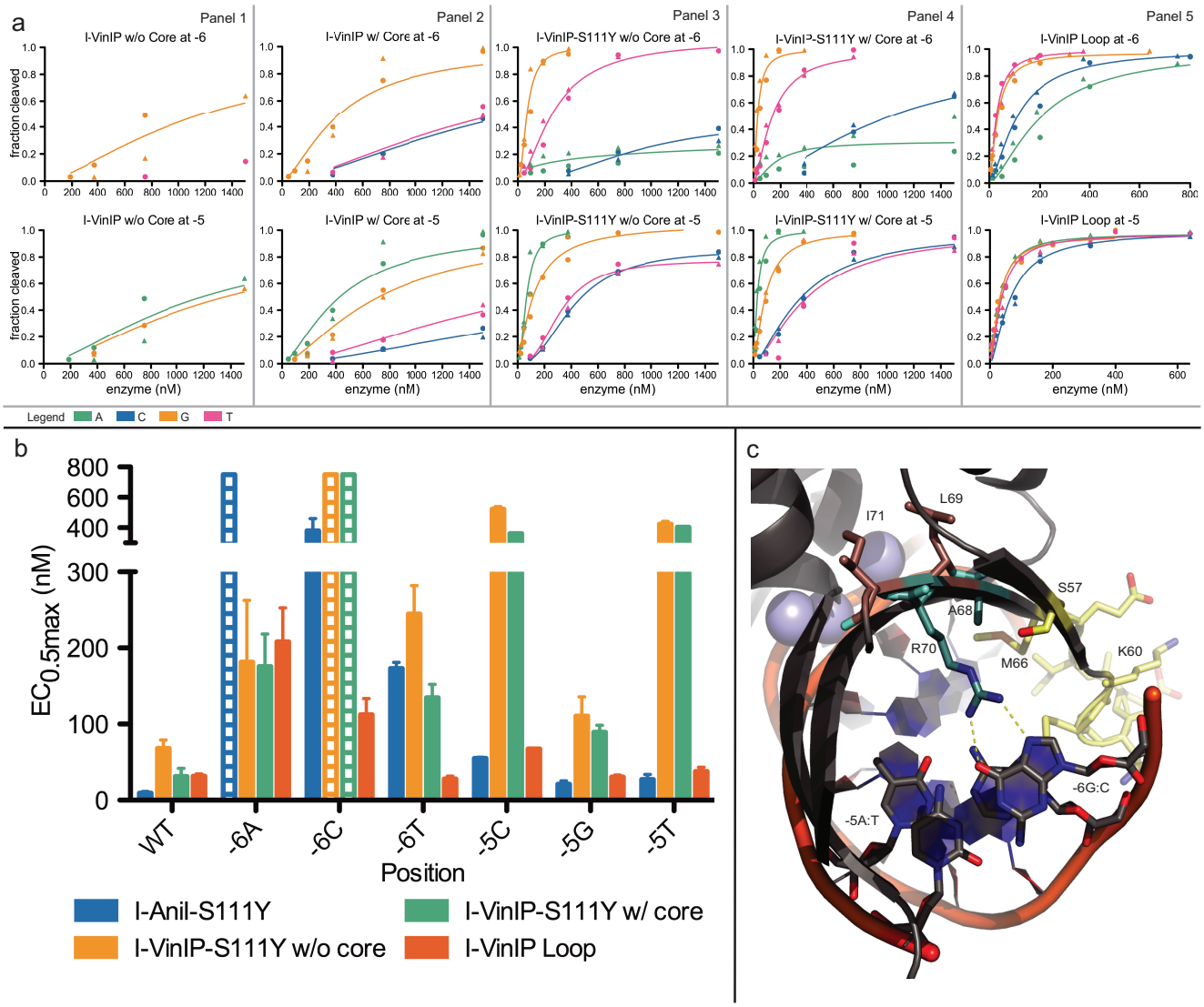


Figure 22. Transfer of homologue-derived core substitutions increases activity. See Table AI.6 for a complete list of the mutations made for every variant protein. **a)** Cleavage profiles for each of the five variants derived from the homologue I-VinIP are given in panels 1 through 5. The upper plot of each panel displays the variants' activity on the singly substituted -6 target site position, while the lower plots show analogous data for position -5. Curves are colored by the substituted base: adenine (green), cytosine (blue), guanine (yellow), or thymine (pink). **b)** The associated $EC_{0.5max}$ values of the five I-VinIP variants representing further quantitative assessment of the data in Figure 3a. Bars with dashed lines indicate substitutions where some cleavage was observed, but $EC_{0.5max}$ was too high (>750nM) to allow accurate quantitative determination. The incorporation of core mutations directly adjacent to the interface mutations surrounding the -6 and -5 positions demonstrated a striking increase in activity. Extending the loop in this region further increased activity and established a preference for -6T over the wild-type -6C. **c)** The relevant substitutions and target site positions are colored by mutation type in this view of the I-AniI crystal structure. The interface substitutions (A68Y and R70K) are shown in cyan, core substitutions (L69F and I71V) are shown in brown, and the positions where additional mutations were incorporated in the I-VinIP Loop variant (Figure 22a, panel 5) are shown in yellow.

(spanning residues 60-66 in I-AniI) and nearby mutation S57G from the I-VinIP homologue to the protein containing both the interface and core changes, as well as the activating S111Y. This hybrid showed an increase in activity for all substitutions at the -5 and -6 positions and a further increase in activity against the -6T (Figure 22a, panel 5, and Figure 22b).

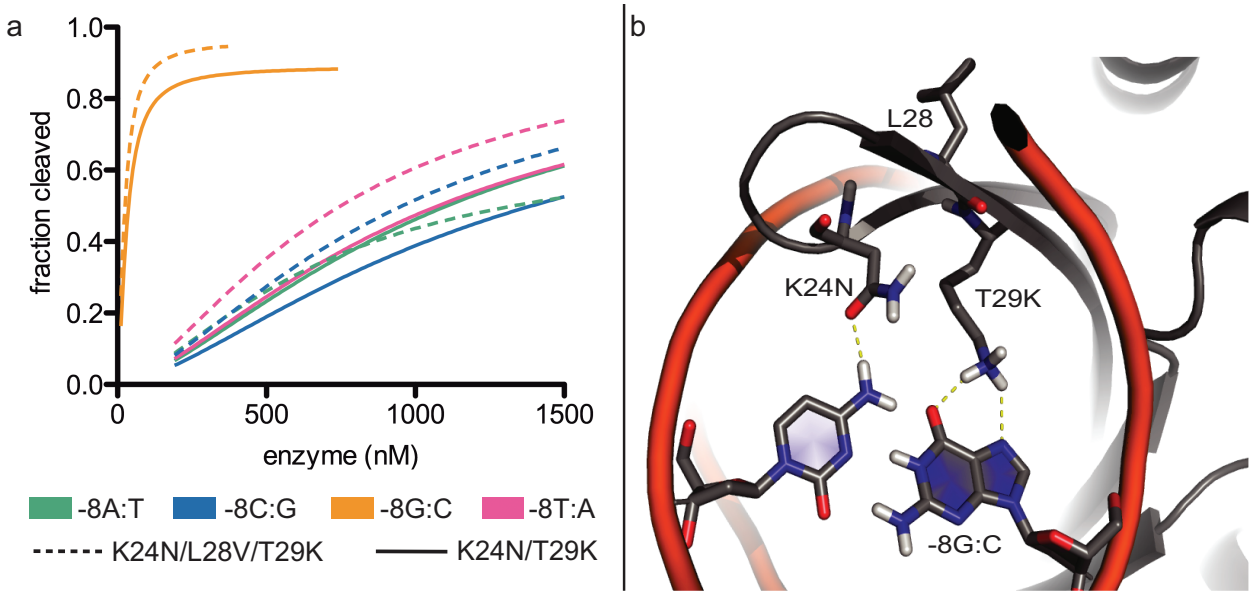


Figure 23. Previously designed variant identified in homologue. See Table AI.6 for a complete list of the mutations made for every variant protein. **a)** Cleavage profiles for the I-AchIP-based variants K24N/T29K and K24N/L28V/T29K on position -8 confirm previously predicted computational mutations for -8G [61]. Introducing the L28V core substitution increased activity slightly at three out of the four possible bases at this position. **b)** The computationally predicted model for these amino acid mutations [61]. N24 forms a hydrogen bond with -8C, while K29 is able to form two hydrogen bonds with -8G.

A second example of core influences on enzyme activity involves transfers from I-AchIP. This homologue contains two interface mutations, K24N and T29K, which were previously predicted computationally by Rosetta [61] and confirmed experimentally to be highly specific for a guanine at position -8 (Figure 23a). Consistent with these results, the predicted homologue target site contains the -8G substitution (Figure 20c). However, in addition to these two mutations, I-AchIP also contains the L28V core substitution. Incorporation of this very conservative mutation resulted in increased activity on three of the four possible nucleotides at position -8 (computationally modeled contacts are visualized in Figure 23b).

5.3 Utility of Homologue Grafting to Endonuclease Engineering

Generation of novel specificities for endonuclease engineering

The use of homing endonucleases for applications such as gene therapy and genome engineering is limited by our ability to target DNA sequences of interest. The endonucleases identified at present recognize only a limited set of DNA targets and thus new approaches are required to access novel DNA target sequences. While computational design has proven successful in generating variants for targets with as many as three base-pair changes [25,50], these methods are still limited to small numbers of substitutions and typically require subsequent, time-consuming, experimental selection [26] to optimize enzyme activity. Information garnered from endonuclease homologues can be used for further design with a two-fold benefit: by gathering a set of previously uncharacterized nucleases with altered binding and cleaving properties to be used as starting scaffolds, and by identifying pockets of mutations that can be grafted to activate the enzyme toward local DNA substitutions.

Grafting of mutations from homologues onto I-AniI resulted in a number of expected and novel specificities (Table 2). The majority of the substitutions identified in the predicted homologue target sites are cleaved, to some extent, by the wild-type enzyme. However, it was hypothesized that amino acid mutations from homologues that neighbor the substituted bases would improve enzymatic activity against these sites. Indeed, several hybrid enzymes were found to have improved activity toward the nucleotide substitutions in their putative target site over the wild-type nucleotide (Figure 20c, Table 2): I-PnoIP N-terminal transfer prefers -5T over -5A, I-VinIP central 4 loop prefers -2C over -2T, and the I-AchIP derived K24N and T29K prefers -8G over -8A.

Conversely, a number of hybrid proteins yielded unexpected specificities given the sequence of their predicted target site. As described in the results section, a single lysine to asparagine mutation at position 200 opens up specificity at position +3 to now allow a cytosine, previously the least favored +3 nucleotide (Figure 21b). Transfer of both the I-AchIP and I-TasIP C-terminal distal loops resulted in activity on the +7A target site that has never before been achieved with either the wild-type I-AniI or any of its designed variants (Figure 21a), and transfer of the I-VinIP interface and core residues neighboring the central four positions allowed

for cleavage of the -6T substitution (Figure 22). These specificity shifts were the most pronounced of all shifts observed for the hybrid enzymes characterized in this study, with all three enzymes tolerating nucleotide substitutions that are not cleaved by wild-type I-AniI. However, all three of these new enzymes have activity against the wild-type I-AniI target indicating that the target site assessments are likely accurate despite the unforeseen specificity changes. For I-AniI, and likely all other endonucleases, neutral drift [69,70] – the accumulation of non-deleterious mutations with adaptive potential – has resulted in the acquisition of new substrate specificities.

Identifying mutations that result in novel specificities, such as the cases discussed above, can directly aid in sequence-specific targeting and modification of disease-causing genes. Table S3 displays five cleavage site sequences in close proximity to chromosomal loci of interest that contain +3C, +7A, and -6T relative to the native I-AniI target site, which were substitutions that could not be cleaved prior to this study. These genes have also recently demonstrated promise as potential therapeutic targets in animal models and perhaps even clinical trials.

Table 3. Sequences of target sites near disease-causing genes that contain the novel +3C, +7A, and -6T specificities. Feasibly targetable cleavage sites in DNA sequences of disease-causing genes were identified computationally according to previously validated methods [50]. For the site sequences given below, bases in lower case lettering differ from the native I-AniI target site. Colored bases indicate the substitutions that could not be cleaved prior to this study: -6T (yellow), +3C (blue), and +7A (green). Human CCR5, a coreceptor for HIV-1, has been successfully targeted and modified by zinc-finger nucleases and is now undergoing clinical trials [14], but there is potential for further modification with homing endonucleases. Nine target site positions differ from that of I-AniI, and we found homologue-derived mutations that can be grafted onto the I-AniI scaffold to target the thymine substitution at position -6 and the adenine at +7. Two target sites found in the murine fumaryl acetoacetate hydrolase gene (FAH, mutations can lead to tyrosinemia are also shown to have sets of substitutions relative to I-AniI that contain +3C, +7A, and -6T. Lastly, recent work has addressed gene delivery in large animal models that shows great potential in preclinical studies, especially canine models. Pyruvate kinase (PK) deficiency in dogs is associated with severe hemolytic anemia that can be treated via allogeneic transplantation, and *in vivo* gene delivery through transduced cells can sufficiently correct the XSCID disease phenotype. Both genes contain targetable sites that can benefit from our novel specificities.

Gene	Target Site
CCR5	ccAGtAtGTTgCcCacaAAA
FAH (Target Site 1)	TGAGccctTTcCcCccTAgc
FAH (Target Site 2)	TGAGtAGcTTTCTCataAgt
PK	cGtGGccaTTgCcCTGgAcA
XSCID	gctGctGcTTctTCTGaccA

The role of core mutations in endonuclease activity

The contribution of surface residues to the activity and specificity of homing endonucleases is generally straightforward, while the role of core mutations on the same has been less well understood. Mutations to the hydrophobic core of I-AniI can yield dramatic effects on activity (94). We find that transfer of a patch of both core and interface mutations from the I-AniI homologue I-VinIP results in greater activity against the -6 position compared with the interface mutations alone (Figure 22). The addition of these core mutations also cause shifts to the specificity at the neighboring DNA bases. In contrast, addition of a single core mutation to the I-AchIP interface mutations K24N and T29K has only a modest effect on activity and almost none on specificity (Figure 23). Further understanding of how core residue changes alter homing endonuclease properties will be important for engineering designs against currently inaccessible target sites.

A new source of data for computational modeling and design

Although gathering information from homing endonuclease homologues can aid in identifying mutations that allow us to access novel targets, more extreme methods may be required to produce enzymes that cleave sequences that differ further from that of the parent endonuclease. The homologues characterized in this paper have evolved to cleave very similar DNA substrates to that of I-AniI, as is apparent when we compare their putative insertion sites (Figure 1c). Transferring pockets of mutations from more divergent homologues is complicated by the challenges of target site identification and identification of residues critical to the switch in specificity. Accessing the vast information in these homologues will require either laborious experimentation or modeling of the homologue in complex with its putative target site.

Computational modeling has yielded accurate predictions for specificity-causing changes, such as the K24N and T29K substitutions that were originally determined by computational prediction to cleave the -8G position [61] and are further validated by their presence in

homologues targeting sites with this substitution. By incorporating information from homologue alignments into standard design protocols it should be possible to engineer enzymes toward more divergent sites. In particular, the modeling of loops [64], how core changes result in protein backbone shifts [65], and sequence preferences of DNA bending [84] are computationally challenging. Improving our understanding of how the amino acid variation alters the specificity and activity of enzymes in relation to these challenges can provide insight into ways to improve computational design methodologies.

Cross-species sequencing provides a useful repository of information on how protein sequence relates to function. For closely related homing endonucleases, it is possible to determine the target DNA of these homologues and to identify amino acid mutations that likely influence target site specificity. We show here that transferring these variable residues to a scaffold protein can be used to determine their effect on both specificity and activity, and that such methods can facilitate the engineering of variant endonucleases with novel target site specificities. This study focuses on aspects of endonuclease structure that are particularly difficult to model, such as flexible loops, the indirect readout of the central four DNA bases, and how core residues influence a protein backbone. Our novel sequence mining approach will likely aid future engineering efforts and provide data that can improve tools for computational prediction – with the benefits not only limited to endonucleases and protein-DNA interactions, but broadly applicable to many enzyme-substrate redesign challenges.

Bibliography

[1] Redondo, P. *et al.* Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature* **456**, 107-11 (2008)
[2] Marcaida, M. J., Munoz, I. G., Blanco, F. J., Prieto, J., & Montoya G. Homing endonucleases: from basis to therapeutic applications. *Cell Mol. Life Sci.* (2009)
[3] Gao, H. *et al.* Heritable targeted mutagenesis in maize using a designed endonuclease. *Plant J.* (2009)
[4] Traver, B. E., Anderson, M. A., & Adelman, Z. N. Homing endonucleases catalyze double-stranded DNA breaks and somatic transgene excision in Aedes aegypti. *Insect Mol. Biol.* **18**, 623-33 (2009)
[5] Hartlerode, A. J. & Scully, R. Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem. J.* **423**, 157-68 (2009)
[6] Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, S., & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636-646 (2010)
[7] Miller, J. C., *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* **27**, 778-785 (2007)
[8] Miller, J. C., *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143-148 (2011)
[9] Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J., Stoddard, B. L. The crystal structure of the TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716-719 (2012)
[10] Gordley, R. M., Gersbach, C. A., & Barbas, C. F. 3rd. Synthesis of programmable integrases. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5053-5058 (2009).
[11] Silva, G., Poirot, L., Galetto, R., Smith, J., Montoya, G., Guchateau, P. & Pâques, F. Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr. Gene Ther.* **11**, 11-27 (2011)
[12] Stoddard, B.L. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure,* **19**, 7-15 (2011)
[13] Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512 (2009)
[14] Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.* **26**, 808-816 (2008)
[15] Remy, S. *et al.* Zinc-finger nucleases: a powerful tool for genetic engineering of animals. *Transgenic Res.* (2009)
[16] Cheng, L., Blazar, B., High, K., & Porteus, M. Zinc fingers hit off target. *Nat. Med.* **17**, 1192-1193 (1011).
[17] Gaj, T., Mercer, A. C., Gersbach, C. A., Gordley, R. M., Barbas, C. F. 3rd. Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 498-503 (2011)
[18] Arnould, S. et al. Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *J. Mol. Biol.* **371**, 49-65 (2007)
[19] Pierce, A. J. & Jasin, M. Measuring recombination proficiency in mouse embryonic stem cells. *Methods Mol. Biol.* **291**, 373-84 (2005)

[20] Geese, W. J., Kwon, Y. K., Wen, X., & Waring, R. B. In vitro analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-AniI. *Eur. J. Biochem.* **270**, 1534-1554 (2003)
[21] Barzel, A., Privman, E., Peeri, M., Naor, A., Schachar, E., Burstein, D., Lazary, R., Gophna, U., Pupko, T., & Kupiec, M. Native homing endonucleases can target conserved genes in humans and animal models. *Nucleic Acids Res.* **39**, 6646-6659 (2011)
[22] Stoddard, B. L. Homing endonuclease structure and function. *Quarterly Reviews in Biophysics* (2005)
[23] Belfort, M. & Perlman, P. S. Mechanisms of intron mobility. *J. Biol. Chem.* **270**, 30237-40 (1995)
[24] Scalley-Kim, M., McConnell-Smith, A. & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**, 1305-1319 (2007)
[25] Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J. Jr., Stoddard, B. L. & Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656-659.
[26] Doyon, J. B., Pattanayak, V., Meyer, C. B. & Liu, D. R. Directed evolution and substrate specificity profiling of homing endonuclease I-SceI. *J. Am. Chem. Soc.* **128**, 2477-84 (2006)
[27] Li, H., Ulge, U. Y., Hovde, B. T., Doyle, L. A., & Monnat, R. J. Jr. Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. *Nucleic Acids Res.* **40**, 2587-2598 (2011)
[28] Speigel, P. C., Chevalier, B., Dussman, D., Turmel, M., Lemieux, C., & Stoddard, B. L. The structure of I-CeuI homing endonuclease: Evolving asymmetric DNA recognition from a symmetric protein scaffold. *Structure* **14**, 869-880 (2006)
[29] Bolduc, J. M. *et al.* Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.* **17**, 2875-2888 (2003)
[30] Takeuchi, R., Certo, M., Caprara, M. G., Scharenberg, A. M., & Stoddard, B. L. Optimization of *in vivo* activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation. *Nucleic Acids Res.* **37**, 877-890 (2008)
[31] Jarjour, J. *et al.* High-resolution profiling of homing endonuclease binding and catalytic specificity using yeast surface display. *Nucleic Acids Res.* **37**, 6871-80 (2009)
[32] McConnel-Smith, A., Takeuchi, R., Pellenz, S., David, L., Maizels, N., Monnat, R. J. Jr., & Stoddard, B. L. Generation of a nicking enzymes that stimulates site-specific gene conversion from the I-AniI LAGLIDADG homing endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5099-104 (2009)
[33] Jencks, W. P. Mechanism of enzyme action. *Annu. Rev. Biochem.* **32**, 639-676 (1963)
[34] Wells, T. N. & Fersht, A. R. Use of binding energy in catalysis measured by mutagenesis of tyrosyl-tRNA synthetase. *Biochemistry* **25**, 1881-1886 (1986)
[35] Fersht, A. R. Relationships between apparent binding energies measured in site-directed mutagenesis experiments and energetics of binding and catalysis. *Biochemistry* **27**, 1577-1580 (1988)
[36] Benkovic, S. J. & Hammes-Schiffer, S. A perspective on enzyme catalysis. *Science* **301**, 1196-1202 (2003)
[37] Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-195 (2008).

[38] Collins, C. H., Yokobayashi, Y., Umeno, D., & Arnold, F. H. Engineering proteins that bind, move, make, and break DNA. *Curr. Opin. Biotechnol.* **14**, 371-378 (2003)
[39] Halford, S. E., Johnson, N. P., & Grinsted, J. The EcoRI restriction endonuclease with bacteriophage lambda DNA. Kinetic studies. *Biochem. J.* **191**, 581-592 (1980)
[40] Silva, G. H., Dalgaard, J. Z., Belfort, M. & Van Roey, P. Crystal structure of the thermostable archaeal intron-endoded endonuclease I-DmoI. *J. Mol. Biol.* **286**, 1123-1136 (1999)
[41] Macaida, M. J. *et al.* Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16888-16893 (2008)
[42] Crothers, D. M. DNA curvature and deformation in protein-DNA complexes: a step in the right direction. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 15163-15165 (1998)
[43] Silva, G. H., Belfort, M., Wende, W. & Pingoud A. From monomeric to homodimeric endonucleases and back: engineering novel specificity of LAGLIDADG enzymes. *J. Mol. Biol.* **361**, 744-754 (2006)
[44] Perrin, A., Buckle, M. & Dujon, B. Asymmetrical recognition and activity of the I-SceI endonuclease on its site and on intron-exon junctions. *EMBO Journal 12*, 2939-47 (1993)
[45] Turmel, M., Mercier, J. P., Cote, V., Otis, C. & Lemieux, C. The site-specific DNA endonuclease encoded by a group I intron in the Chlamydomonas pallidostigmatica chloroplast small subunit rRNA gene introduces a single-strand break at low concentrations of Mg2+. *Nucleic Acids Res* **23**, 2519-25 (1995)
[46] Kalodimo, C. G., Boelens, R. & Kaptein, R. A residue-specific view of the association and dissociation pathway in protein-DNA recognition. *Nat. Struct. Biol.* **9**, 193-197 (2002)
[47] Havranek, J. J. & Baker, D. Motif-directed flexible backbone design of functional interactions. *Protein Sci.* **18**, 1293-205 (2009)
[48] Dunbrack, R. L. Jr. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661-8 (1997)
[49] Havranek, J. J. Duarte, C. M. & Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.* **344**, 59-70 (2004)
[50] Ashworth, J., Taylor, G. K., Havranek, J. J., Quadri, S. A., Stoddard, B. L., & Baker, D. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38**, 5601-5608 (2010)
[51] Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12**, 963-72 (2003)
[52] Wang, C., Bradley, P. & Baker, D. Protein-protein docking with backbone flexibility. *J. Mol. Biol.* **373**, 503-19 (2007)
[53] Yanover, C. & Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* **39**, 4564-4576 (2011)
[54] Morozov, A. V., Havranek, J. J., Baker, D. & Siggia, E. D. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.* **33**, 5781-5798 (2005)
[55] Ashworth, J. & Baker, D. Assessment of the optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* **37**, e73 (2009)
[56] Windbichler, N. *et al.* A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature* **473**, 212-215 (2011)

[57] Chames, P., Epinat, J. C., Guillier, S., Patin, A., Lacroix, E. & Pâques, F. *In vivo* selection of engineered homing endonucleases using double-strand break induced homologous recombination. *Nucleic Acids Res.* **33**, e178 (2005)
[58] Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J. Jr. & Stoddard, B. L. Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell* **10**, 895-905 (2002)
[59] Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. Computational method to reduce the search space for directed protein evolution. *PNAS* **98**, 3778-3783 (2001)
[60] Leaver-Fay, A. *et al.* Rosetta3: an object-oriented software suite for simulation and design of macromolecules. *Methods Enzymol.* **487**, 545-574 (2011)
[61] Thyme, S. B., Jarjour, J., Takeuchi, R. Havranek, J. J., Ashworth, J., Scharenberg, A. M., Stoddard, B. L. & Baker, D. Exploitation of binding energy for catalysis and design. *Nature* **461**, 1300-1304 (2009)
[62] Ulge, U. Y., Baker, D. A. and Monnat, R. J. Jr. Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. *Nucleic Acids Res.* **39**, 4330-4339 (2011)
[63] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000)
[64] Murphy, P. M., Bolduc, J. M., Gallaher, J. L., Stoddard, B. L. & Baker, D. Alteration of enzyme specificity by computational loop modeling and design. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9215-9220 (2009)
[65] Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838 (2011)
[66] Lazaridis T & Karplus M. Effective energy function for proteins in solution. *Proteins* **35**, 133-152 (1999)
[67] Frankel, A. D. & Kim, P. S. Modular structure of transcription factors: implications for gene regulation. *Cell* **165**, 717-719 (1991)
[68] Szeto, M. D., Boissel, S. J., Baker, D. & Thyme, S. B. Mining endonuclease cleavage determinants in genomic sequence data. *J. Biol. Chem.* **286**, 32617-32627 (2011)
[69] Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.*, **1**, 67-78 (2007)
[70] Bloom, J. D, Romero, P. A., Lu, Z. & Arnold, F. H. Neutral drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direc*t, **2**, 17 (2007)
[71] Lin, J. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* **37**, 145-151 (1991)
[72] Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z. & Players F. Predicting protein structures with a multiplayer online game. *Nature* **466**, 756-760 (2010)
[73] Fleishman, S. J., Khare, S. D., Koga, N. & Baker, D. Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **20**, 753-757 (2011)
[74] Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica* **D66**,12-21 (2010)
[75] Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375-W383 (2007)

[76] Chen, V. B. *et al.* KiNG (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program. *Protein Science* **18**, 2403-2409 (2009)
[77] Matthews, B.W. Protein-DNA interaction. No code for recognition. *Nature* **335**, 294-295 (1988)
[78] Pabo, C.O. & Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597-624 (200)
[79] Temiz, N.A. & Camacho, C.J. Experimentally based contact energies decode interactions responsible for protein–DNA affinity and the role of molecular waters at the binding interface, *Nucleic Acids Res.* **37**, 4076-4088 (2009)
[80] Alibes, A., Serrano, L. & Nadra, A. D. Structure-based DNA-binding prediction and specificity. *Methods Mol. Biol.* **649**, 77-88 (2010)
[81] Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* **9**, 435-442 (2011)
[82] Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant sidechain prediction. *J. Mol. Biol.* **380**, 742-756 (2008)
[83] Steffen, N. R., Murphy, S. D., Tolleri, L., Hatfield, G. W. & Lathrop R. H. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics.***18**, S22–S30 (2002)
[84] Becker, N. B., Wolff, L., and Everaers, R. Indirect readout: detection of optimized sequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.*, **34**, 5638-5649 (2006)
[85] Smith, C. A. & Kortemme, T. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. *PLoS One* **6**, e20451 (2011)
[86] Fu, X., Apgar, J. R. & Keating, A. E. Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-XL. *J. Mol. Biol.* **371**, 1099-1117 (2007)
[87] Kono, H. & Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **35**, 114-131 (1999)
[88] Chevalier, B. & Stoddard, B.L. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res.* **29**, 3757-3774 (2001)
[89] Chevalier, B., Monnat, R.J. Jr., & Stoddard, B.L. in *Homing Endonucleases and Inteins* (Belfort, M., Derbyshire, V., Wood, D. and Stoddard, B.L., eds) pp. 33-47, Springer Verlag, Berlin/Heidelberg (2005)
[90] Chevalier, B., Turmel, M., Lemieux, C., Monnat, R.J. Jr., & Stoddard, B.L. Metal-dependent DNA cleavage mechanism of the I-CreI LAGLIDADG homing endonculease. *J. Mol. Biol*. **329**, 253-269 (2003)
[91] Moure, C.M., Gimble, F.S., & Quiocho, F.A. The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. *J. Mol. Biol.* **334**, 685–695 (2003)
[92] Gimble, F.S. Degeneration of a homing endonuclease and its target site in a wild yeast strain. *Nucleic Acids Res.* **29**, 4215-4223 (2001)
[93] Ho, Y., Kim, S.J., & Waring, R.B. A protein encoded by a group I intron in Aspergillus nidulans directly assists in RNA splicing and is a DNA endonuclease. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8994–8999 (1997)

[94] Longo, A., Leonard, C.W., Bassi, G.S., Berndt, D., Krahn, J., Hall, T.M., & Weeks, K.M. Evolution from DNA to RNA recognition by the bI3 LAGLIDAG maturase. *Nat. Struct. Mol. Biol.* **12**, 779-87 (2005)
[95] Grishin, A., Fonfara, I., Alexeevski, A., Spirin, S., Zanegina, O., Karyagina, A., Alexeyevsky, D., & Wende, W. Identification of conserved features of LAGLIDADG homing endonucleases. *J. Bioinform. Comput. Biol*. **8**, 453-469 (2010)

Appendix I: Supplemental Figures and Tables

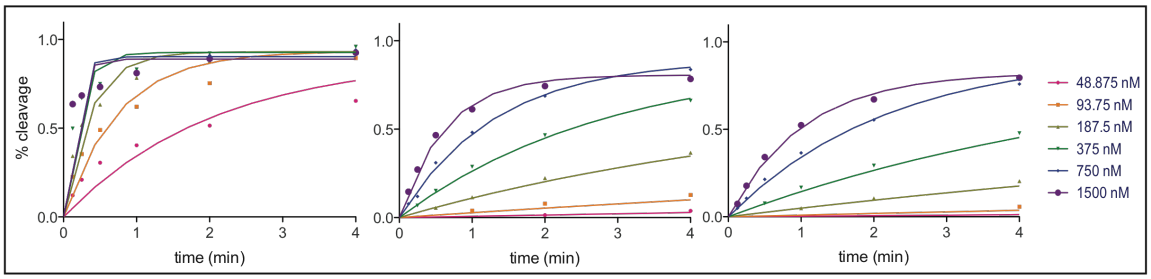
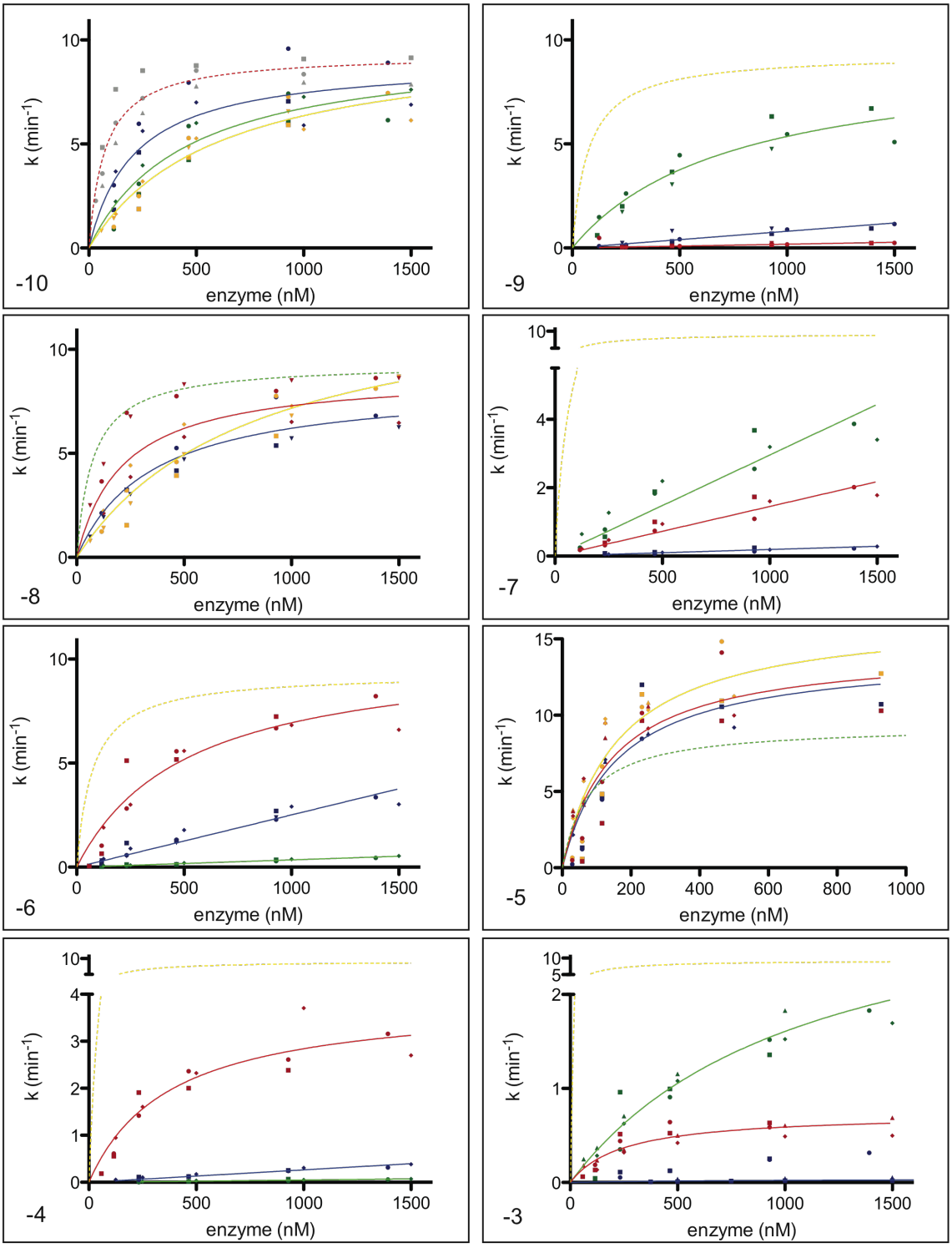
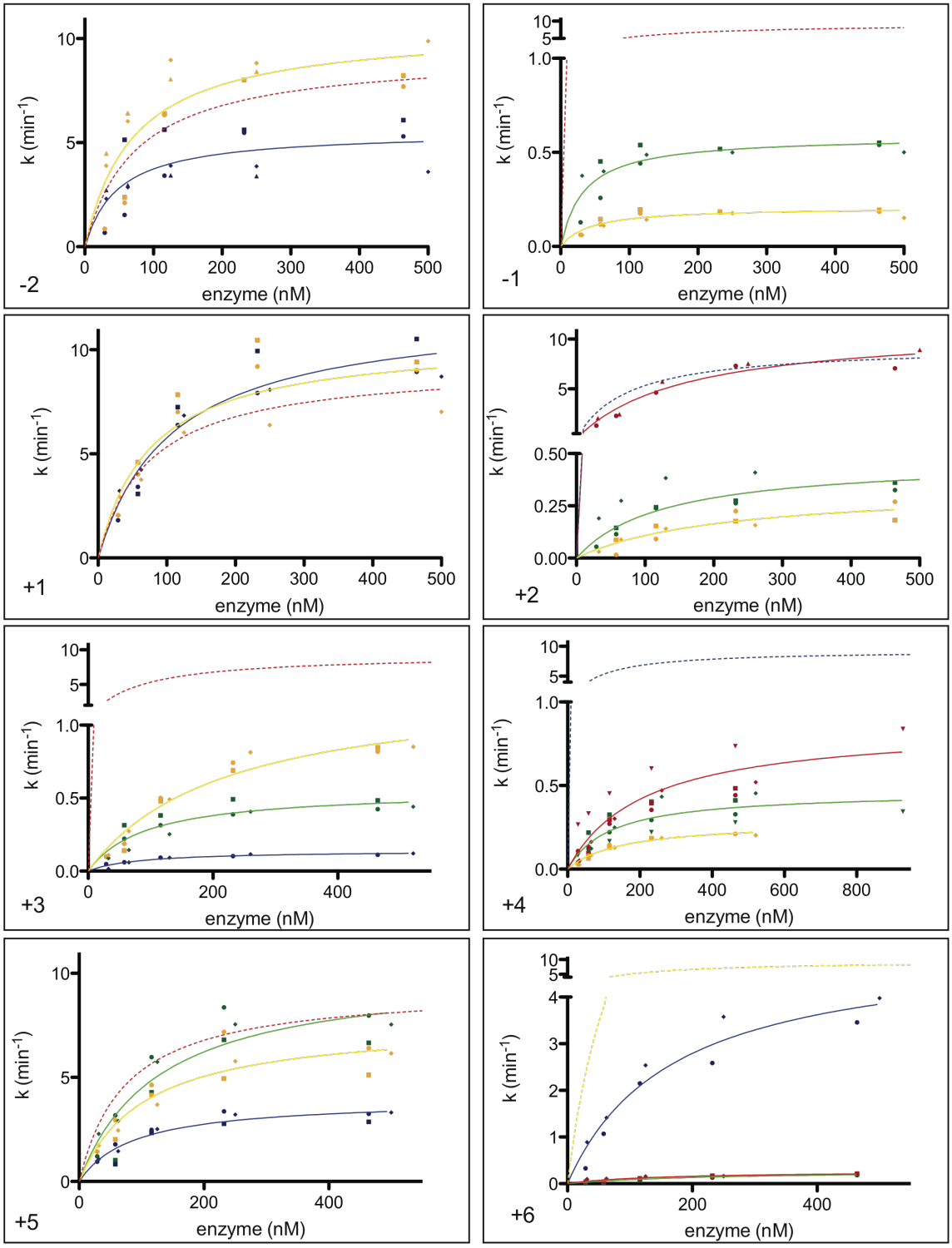


Figure AI.1. Three single base-pair variants in the central four base pairs from the wild-type target site display behavior that cannot be fit with a single-exponential (-2A, -1C, and +1A) when kinetics are collected with wild-type and Y2 I-AniI (data not shown). (Left plot) wild-type enzyme (not Y2) cleaving the LIB4 target site – a variant site with increased binding affinity that includes the +1A mutation (as well as the +8T mutation)[a] – cannot be fit with a single-exponential. (Middle plot) the original wild-type enzyme cleaving the LIB4 target site with an additional -9A substitution can be fit with a single-exponential. The -9A target site has significantly increased K_M* compared to the wild-type target site. (Right plot) the -9C:G designed enzyme cleaving the LIB4 target site (not containing the intended -9C substitution) can be fit with a single-exponential. The deviation from simple exponential kinetics evidently is a function of the interaction between the enzyme and target site. A comparison of the left and middle plots indicate that the deviation is not inherent in the enzyme, and comparison of the left and right plots indicate that it is not inherent in the LIB4 target site. The deviations from single exponential kinetics could result from a non-productive mode of binding or a slow DNA conformational change, for several of the substituted central four substrates (-2A, -1C, +1A), that is only evident when the target site is tightly bound by the enzyme (both the middle and right plots are under conditions of suboptimal binding).

[a] Scalley-Kim, M., McConnell-Smith, A., & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**, 1305-1319 (2007)


-10
-9
-8
-7
-6
-5
-4
-3
k (min-1)
enzyme (nM)


k (min-1)
enzyme (nM)
-2
-1
+1
+2
+3
+4
+5
+6

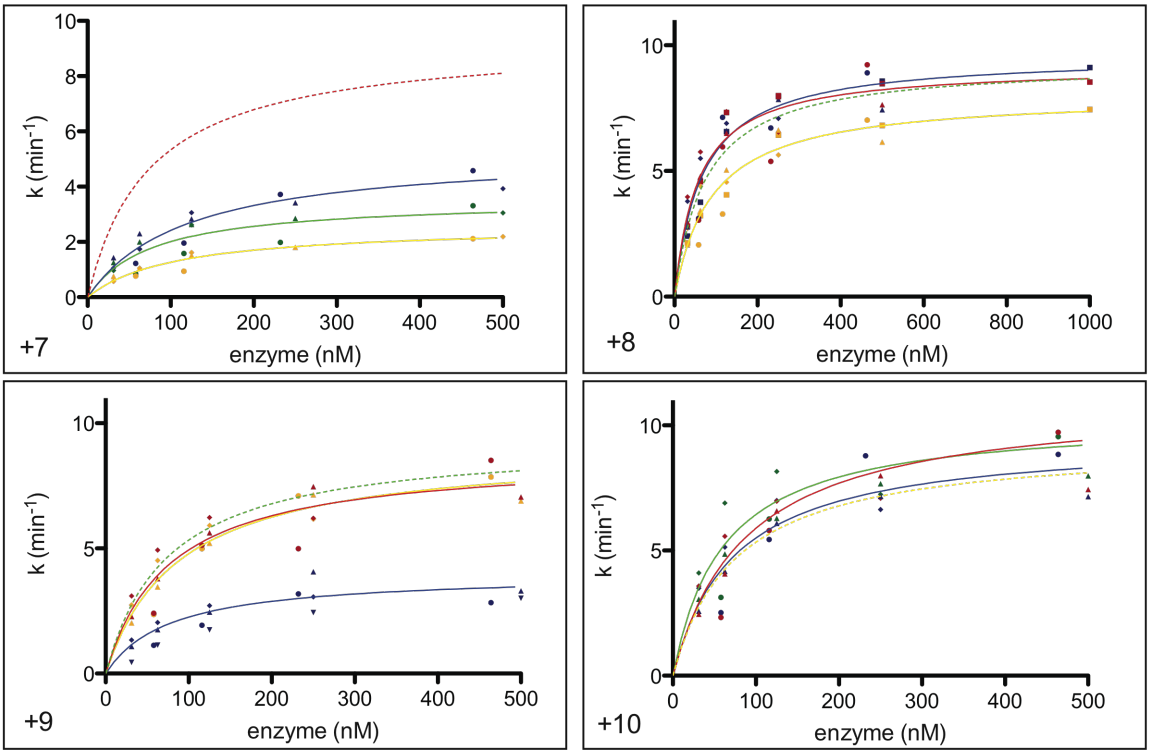


Figure AI.2. Michaelis-Menten plots for each of the 60 singly-substituted target sites in order from left to right across the target site. The data is grouped by position and colored by nucleotide (A=green, C=blue, G=yellow, T=red). The Michaelis-Menten curve fit for the wild-type target site is included on every graph as a dashed line with the color of the wild-type base-pair at that position; the actual data points for the wild-type substrate are shown only in the -10 position panel (grey).

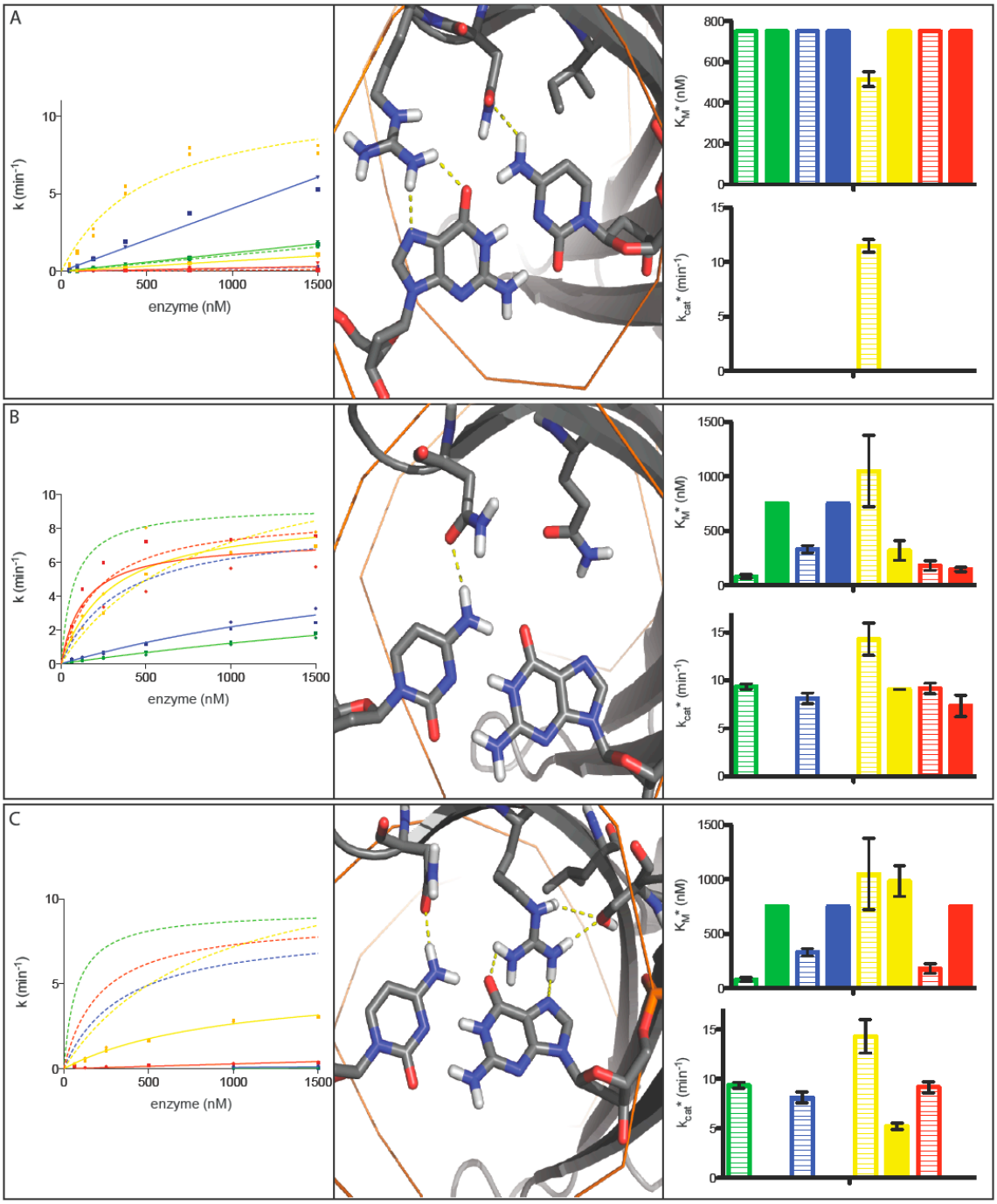
Table AI.1. Kinetic parameters for singly-substituted sites. *

position	base-pair	k_{cat}* / K_M*	k_{cat}* (min^{-1})	K_M* (nM)	Relative K_A
wild-type	wild-type	0.13 ± 0.04	9.4 ± 0.3	81 ± 17	1.0
-10T	A	0.021 ± 0.004	10.0 ± 0.5	505 ± 120	-0.073 ± 0.030
	C	0.052 ± 0.014	9.4 ± 1.2	219 ± 70	0.38 ± 0.03
	G	0.017 ± 0.002	11.4 ± 1.2	748 ± 157	-0.054 ± 0.040
-9G	A	0.013 ± 0.003	10.5 ± 1.9	946 ± 294	-0.23 ± 0.03
	C	0.0010 ± 0.0001	-	>750	-0.16 ± 0.06
	T	0.00020 ± 0.00002	-	>750	-0.28 ± 0.03
-8A	C	0.025 ± 0.002	8.1 ± 0.6	330 ± 34	0.051 ± 0.048
	G	0.017 ± 0.004	14.3 ± 1.7	1048 ± 328	-0.078 ± 0.018
	T	0.058 ± 0.014	9.2 ± 0.6	180 ± 49	-0.11 ± 0.02
-7G	A	0.003 ± 0.0004	-	>750	-0.13 ± 0.04
	C	0.00020 ± 0.00003	-	>750	-0.17 ± 0.06
	T	0.0016 ± 0.0002	-	>750	-0.17 ± 0.02
-6G	A	0.00035 ± 0.00002	-	>750	-0.22 ± 0.02
	C	0.0030 ±0.0001	-	>750	-0.071 ± 0.045
	T	0.020 ± 0.002	11.3 ± 1.1	583 ± 102	-0.087 ± 0.010

-5A	C	0.104 ± 0.009	13.0 ± 0.8	129 ± 19	0.36 ± 0.06
	G	0.130 ± 0.032	15.6 ± 1.1	144 ± 49	0.41 ± 0.05
	T	0.13 ± 0.04	14.3 ± 1.5	153 ± 60	0.46 ± 0.09
-4G	A	0.000053 ± 0.000009	-	>750	0.26 ± 0.02
	C	0.00027 ± 0.00001	-	>750	0.098 ± 0.05
	T	0.0104 ± 0.0007	3.8 ± 0.2	374 ± 40	0.15 ± 0.04
-3G	A	0.003 ± 0.001	3.4 ± 0.6	1134 ± 317	0.20 ± 0.06
	C	0.00007 ± 0.00003	0.046 ± 0.008	874 ± 204	0.30 ± 0.04
	T	0.0028 ± 0.0003	0.78 ± 0.06	281 ± 29	0.030 ± 0.033
-2T	A	-	-	-	1.84 ± 0.25
	C	0.28 ± 0.11	5.4 ± 1.0	53 ± 39	1.12 ± 0.05
	G	0.17 ± 0.05	10.6 ± 0.3	86 ± 26	2.14 ± 0.12
-1T	A	0.03 ± 0.01	0.58 ± 0.04	36 ± 23	1.66 ± 0.05
	C	-	-	-	0.60 ± 0.05
	G	0.006 ± 0.002	0.20 ± 0.009	39 ± 10	2.05 ± 0.09
+1T	A	-	-	-	1.76 ± 0.12
	C	0.12 ± 0.01	12.1 ± 1.3	109 ± 19	1.28 ± 0.13
	G	0.15 ± 0.01	10.6 ± 1.3	74 ± 12	1.67 ± 0.36
+2C	A	0.03 ± 0.02	0.38 ± 0.05	81 ± 39	1.49 ± 0.15
	G	0.003 ± 0.001	0.23 ± 0.02	97 ± 31	1.89 ± 0.35
	T	0.071 ± 0.002	11.2 ± 1.0	157 ± 11	0.34 ± 0.06
+3T	A	0.0030 ± 0.0004	0.56 ± 0.03	104 ± 42	1.63 ± 0.12
	C	0.002 ± 0.001	0.14 ± 0.02	84 ± 34	0.83 ± 0.08
	G	0.0060 ± 0.0003	1.3 ± 0.03	222 ± 17	1.49 ± 0.03
+4C	A	0.004 ± 0.001	0.51 ± 0.07	150 ± 36	1.02 ± 0.06
	G	0.0020 ± 0.0002	0.32 ± 0.04	185 ± 39	1.35 ± 0.13
	T	0.005 ± 0.001	0.75 ± 0.06	171 ± 27	0.77 ± 0.08
+5T	A	0.080 ± 0.013	10.4 ± 0.4	139 ± 28	1.87 ± 0.31
	C	0.045 ± 0.006	4.0 ± 0.09	92 ± 10	1.07 ± 0.12
	G	0.074 ± 0.007	7.7 ± 0.7	106 ± 12	1.52 ± 0.16
+6G	A	0.00100 ± 0.00007	0.45 ± 0.002	614 ± 51	1.19 ± 0.06
	C	0.031 ± 0.006	5.1 ± 0.2	170 ± 24	1.14 ± 0.12
	T	0.00120 ± 0.00006	0.34 ± 0.034	282 ± 41	0.74 ± 0.14
+7T	A	0.047 ± 0.016	4.5 ± 0.6	157 ± 91	1.25 ± 0.13
	C	0.052 ± 0.014	5.8 ± 1.0	147 ± 68	0.94 ± 0.02
	G	0.026 ± 0.006	2.9 ± 0.3	134 ± 47	1.68 ± 0.18
+8A	C	0.16 ± 0.03	9.5 ± 0.5	68± 15	1.18 ± 0.06
	G	0.10 ± 0.03	8.3 ± 1.0	118 ± 51	1.54 ± 0.25
	T	0.18 ± 0.05	9.5 ± 0.8	74 ± 28	1.38 ± 0.11
+9A	C	0.046 ± 0.01	3.9 ± 0.2	96 ± 18	1.42 ± 0.26
	G	0.11 ± 0.03	9.1 ± 1.0	96 ± 30	1.29 ± 0.14
	T	0.13 ± 0.04	9.4 ± 1.3	108 ± 53	1.41 ± 0.22
+10A	A	0.20 ± 0.07	10.6 ± 1.6	71 ± 31	1.04 ± 0.11
	C	0.15 ± 0.04	9.8 ± 1.5	85 ± 37	1.06 ± 0.13

	T	0.15 ± 0.04	11.0 ± 2.1	90 ± 39	1.00 ± 0.10

*All values are the mean ± the standard error from mean (SEM) from 2-4 independent reaction velocity versus enzyme concentration experiments.



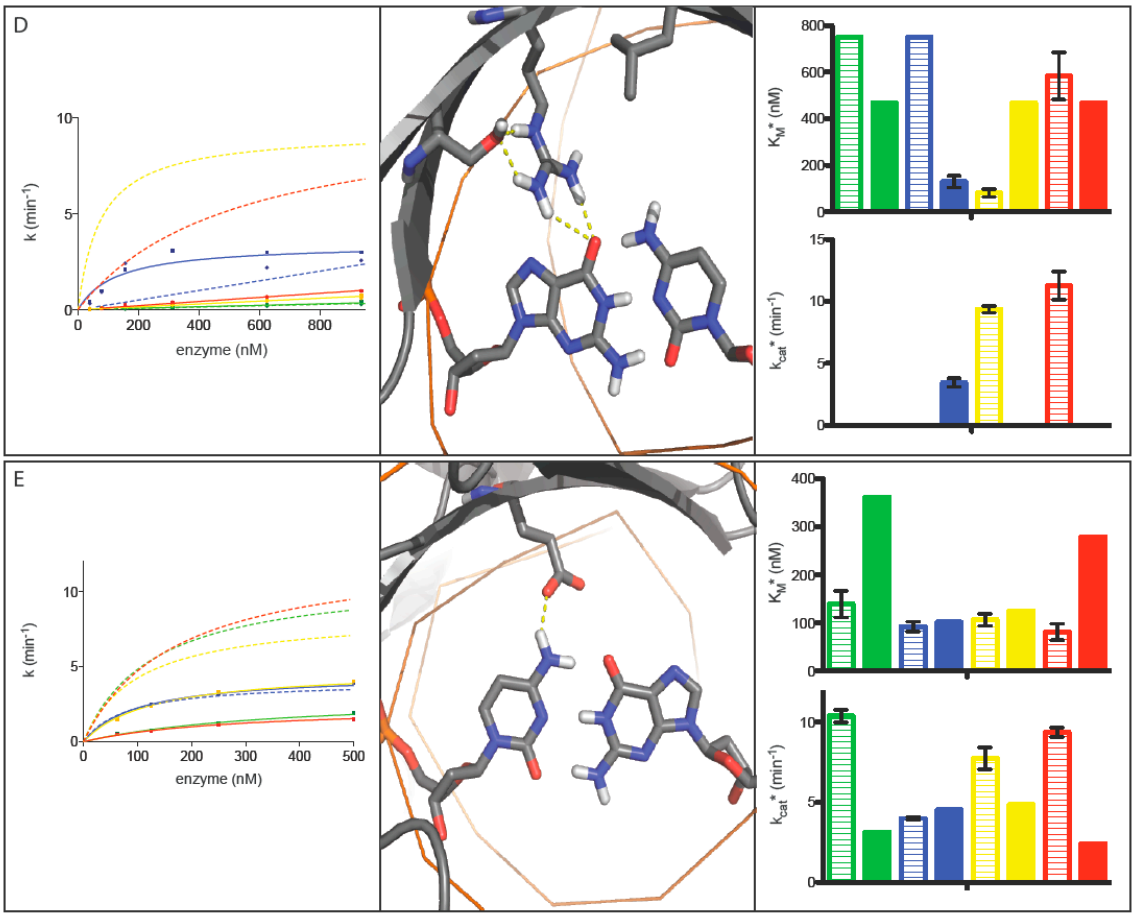


Figure AI.3. The color scheme throughout the figure is A=green, C=blue, G=yellow,T=red, and error bars in right panels are standard errors from the mean (SEM).
a) Design for -9G:C to -9C:G substitution (K24N, G25R, T29I). (Middle panel) The designed residue R25 makes hydrogen bonds to -9G, and N24 makes a hydrogen bond to -9C. I29 removes contacts with the wild-type -9G:C and provide space for the new contacts provided by R25 and N24. (Left panel, right panel) this enzyme overall binds significantly more poorly than wild-type enzyme, in part because it is missing the Y2 mutations that enhance binding (due to expression issues with the Y2 version of this design). However, the enzyme is specific for the designed site, which was not tolerated at all by the wild-type enzyme, and the specificity appears to be derived from modulation of K_M*. **b)** Design B for -8A:T to -8G:C substitution (K24N, T29Q). (Middle panel) The designed residue N24 makes hydrogen bonds to -8C (as in the other two designs for this same base-pair). (Left panel, right panel) this design is equally specific for the -8T:A and the target -8G:C, however it has significantly increased K_M* for the other two substitutions (the altered specificity is achieved through K_M*, not k_{cat}*). **c)** Design C for -8A:T to -8G:C substitution (K24N, T29R, E31L, R72S). (Middle panel) The designed residue R29 makes hydrogen bonds to -8G, and N24 makes hydrogen bonds to -8C (as in the other two designs for this same base-pair). S72 is positioning R29 and L31 is necessary to allow space for R29 to contact -8G (additional, it is likely electrostatic influences of E31 would have disrupted the R29 contact). (Left panel, right panel) this enzyme overall binds significantly more poorly than wild-type enzyme. However, the enzyme is very specific for the designed site over the other three possible target

sites (significantly more specific than the wild-type enzyme at this position) and the specificity appears to be derived from modulation of K_M*. **d)** Design for -6G:C to -6C:G substitution (T29S, E31R, R70L). (Middle panel) The designed residue R31 makes direct hydrogen bonds to -6G, and its position is stabilized by hydrogen bonds with S29. The L70 mutation is necessary to remove direct interactions with the wild-type base-pair at this position. (Left panel, right panel) this enzyme overall binds significantly more poorly than wild-type enzyme. However, the enzyme is very specific for the designed site over the other three possible target sites. **e)** Design for +5T:A to +5C:G substitution (D168E). (Middle panel) the designed residue E168 makes a direct hydrogen bond with +5C. (Left panel, right panel) the shifts in activity for this enzyme results mainly from modulation of k_{cat}* and a switching of the order of k_{cat}* preferences at this position, resulting in enhanced specificity over the wild-type enzyme.

Table AI.2. Kinetic parameters for designed enzymes[b]

Design	Mutations	Base-pair[c]	WT k_{cat}* (min^{-1})	WT K_M* (nM)	WT k_{cat}* / K_M*	Variant k_{cat}* (min^{-1})	Variant K_M* (nM)	Variant k_{cat}* / K_M*	Exp. Spec.	Pred. Spec.
-9G:C to -9C:G[d,e]	K24N, G25R, T29I, F80K[f], L233K[f]	A	-	>750	0.0010 ± 0.00002	-	>750	0.00100 ± 0.00006	0.53	0.28
		C	-	>750	0.00020 ± 0.000005	-	>750	0.0040 ± 0.0002		
		G	11.5 ± 0.6	516 ± 37	0.022 ± 0.0004	-	>750	0.00070 ± 0.00005		
		T	-	>750	0.000071 ± 0.000001	-	>750	0.0002 ± 0.0001		
-8A:T to -8G:C_A	K24N, T29K	A	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	8.3 ± 0.1	2237 ± 17	0.00400 ± 0.00008	0.87	0.96
		C	8.1 ± 0.6	330 ± 34	0.025 ± 0.002	6.0 ± 0.9	1821 ± 466	0.0030 ± 0.0004		
		G	14.3 ± 1.7	1048 ± 328	0.017 ± 0.004	9.9 ± 0.01	26 ± 4	0.40 ± 0.06		
		T	9.2 ± 0.6	180 ± 49	0.058 ± 0.014	9.7 ± 0.9	1535 ± 117	0.0060 ± 0.001		
-8A:T to -8G:C_B	K24N, T29Q	A	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	-	>750	0.00100 ± 0.00009	0.37	0.42
		C	8.1 ± 0.6	330 ± 34	0.025 ± 0.002	-	>750	0.0020 ± 0.0002		
		G	14.3 ± 1.7	1048 ± 328	0.017 ± 0.004	9.06 ± 0.01	320 ± 90	0.031 ± 0.009		
		T	9.2 ± 0.6	180 ± 49	0.058 ± 0.014	7.4 ± 1.1	144 ± 20	0.053 ± 0.015		
-8A:T to -8G:C_C[d]	K24N, T29R, R72S, E31L, F80K[f],	A	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	-	>750	0.000025 ± 0.0000005	0.82	0.58
		C	8.1 ± 0.6	330 ± 34	0.025 ± 0.002	-	>750	0.000073 ± 0.000047		
		G	14.3 ± 1.7	1048 ± 328	0.017 ±	5.2 ± 0.3	983 ± 141	0.0050 ±		

[b] All values are the mean ± the standard error from mean (SEM) from 2-4 independent reaction velocity versus enzyme concentration experiments.
[c] The wild-type base-pair is highlighted in green and the target base-pair highlighted in cyan.
[d] These enzymes were impure and the concentration of enzyme was calculated using the SDS-PAGE method described in the supplementary material.
[e] This variant was tested in the wild-type background with the LIB4 target site due to expression issues with the Y2 version. The enhanced binding of the LIB4 target site (Scalley-Kim, M., McConnell-Smith, A., & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**, 1305-1319 (2007)) compensated for the binding loss associated with missing the Y2 mutations.
[f] These mutations were included for solubility. Whether or not an enzyme contains these mutations was a function of date of gene construction.

	L233K[f]	G	14.3 ± 1.7	1048 ± 328	0.004	5.2 ± 0.3	983 ± 141	0.0004		
		T	9.2 ± 0.6	180 ± 49	0.058 ± 0.014	-	>750	0.00028 ± 0.00001		
-6G:C to -6C:G[d]	T29S, E31R, R70L, L233K[f]	A	-	>750	0.00035 ± 0.00002	-	>469	0.00040 ± 0.00008	0.79	0.94
		C	-	>750	0.0030 ±0.0001	129 ± 25	3.5 ± 0.4	0.027 ± 0.003		
		G	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	-	>469	0.00100 ± 0.00008		
		T	11.3 ± 1.1	583 ± 102	0.020 ± 0.002	-	>469	0.00100 ± 0.000002		
-3G:C to -3C:G	Y18W, E35K, R61Q, L233K[f]	A	3.4 ± 0.6	1134 ± 317	0.0030 ± 0.0004	6.3 ± 1.2	3070 ± 732	0.00200 ± 0.00008	0.61	0.81
		C	0.046 ± 0.008	874 ± 204	0.00007 ± 0.00003	2.1 ± 0.1	173 ± 5	0.0120 ± 0.0001		
		G	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	6.3 ± 0.5	6261 ± 638	0.00100 ± 0.00003		
		T	0.78 ± 0.06	281 ± 29	0.0028 ± 0.0003	0.46 ± 0.12	730 ± 259	0.00070 ± 0.00007		
+5T:A to +5C:G	D168E, F80K[f], L233K[f]	A	10.4 ± 0.4	139 ± 28	0.080 ± 0.013	3.3 ± 0.2	644 ± 283	0.006 ± 0.002	0.41	0.22
		C	4.0 ± 0.09	92 ± 10	0.045 ± 0.006	4.6 ± 0.1	159 ± 58	0.033 ± 0.012		
		G	7.7 ± 0.7	106 ± 12	0.074 ± 0.007	4.8 ± 0.1	283 ± 158	0.025 ± 0.014		
		T	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	2.7 ± 0.3	779 ± 500	0.005 ± 0.003		
+8A:T to +8C:G	L156Q, I164R, T189S, F80K[f], L233K[f]	A	9.4 ± 0.3	81 ± 17	0.13 ± 0.04	1.1 ± 0.1	169 ± 22	0.007 ± 0.001	0.57	0.83
		C	9.5 ± 0.5	68± 15	0.16 ± 0.03	9.9 ± 0.7	83 ± 3	0.118 ± 0.004		
		G	8.3 ± 1.0	118 ± 51	0.10 ± 0.03	0.89 ± 0.18	78 ± 52	0.018 ± 0.009		
		T	9.5 ± 0.8	74 ± 28	0.18 ± 0.05	1.3 ± 0.2	147 ± 32	0.0090 ± 0.0007		

Table AI.3. Percent recoveries for iterations of ROSETTA energy function optimization.

	Unweighted Recovery						Weighted Recovery					
Motif Weight	**NA***	**0.00**	**-1.25**	**-2.50**	**-5.00**	**-10.0**	**NA***	**0.00**	**-1.25**	**-2.50**	**-5.00**	**-10.0**
Standard Training	28.25± 0.02	28.92 ±0.06	34.16± 0.14	36.17±0. 23	35.40 ±0.08	33.75 ±0.19	29.64 ±0.06	29.96 ±0.00	32.28± 0.02	33.33± 0.12	31.11 ±0.13	28.83± 0.10
Standard Test	29.54± 0.13	30.93 ±0.22	35.09± 0.20	36.24±0. 29	35.71 ±0.22	33.55 ±0.13	31.55 ±0.26	32.24 ±0.22	33.25± 0.25	33.27± 0.19	30.81 ±0.00	27.93± 0.43
Stringent HBonds Training	32.0	34.3	38.4	39.8	39.6	36.9	32.7	33.7	35.1	35.3	33.4	29.8
Stringent HBonds Test	33.2	35.8	39.5	40.4	40.0	37.0	32.9	35.3	36.0	35.8	33.8	29.7
Phos. Desolvation Training	35.2	37.2	42.1	42.6	41.8	38.7	34.4	35.9	38.2	37.7	34.9	31.0
Phos. Desolvation Test	38.0	40.3	43.5	44.2	42.4	39.3	39.2	40.5	41.4	39.9	35.9	32.2
Electrostatics Training	40.0	41.8	46.3	46.0	44.1	41.0	37.4	38.4	41.4	39.7	35.8	31.7
Electrostatics Test	42.6	45.1	46.8	47.2	44.7	40.9	41.5	43.1	42.3	41.8	37.2	32.8
LK_Ball Training	39.1	41.4	44.1	45.1	44.0	40.8	39.0	39.5	40.8	40.0	36.7	32.4
LK_Ball Test	40.1	42.5	45.1	45.9	44.0	40.2	40.0	42.3	43.3	42.6	38.2	33.0
Multi-Desolvation Training	41.9	44.2	46.7	47.4	45.4	42.3	39.3	40.8	41.5	41.5	37.3	33.0
Multi-Desolvation Test	43.2	45.3	47.6	47.3	45.2	41.1	41.0	42.4	45.0	42.0	38.2	32.8
Attractive (fa_atr) Training	43.5	46.0	47.3	46.2	44.7	42.4	39.2	40.7	40.6	38.5	35.7	32.8
Attractive (fa_atr)	45.0	46.2	47.8	47.6	45.6	41.4	42.3	41.8	43.9	41.5	38.4	33.1

Test												
Lysine Charge Training Lysine Charge Test	44.0 45.2	46.0 47.9	47.8 48.4	47.1 48.2	45.6 46.0	42.9 41.8	38.7 41.3	40.7 43.4	40.6 43.6	39.2 41.2	37.1 37.7	33.4 32.5
Reference Energies Training Reference Energiest Test	45.2 46.8	49.2 49.3	50.0 50.7	48.5 50.0	46.8 46.4	43.9 42.3	41.0 43.2	43.3 44.8	42.9 45.1	39.6 43.0	37.3 38.5	33.7 33.6

*NA = No motif rotamers added

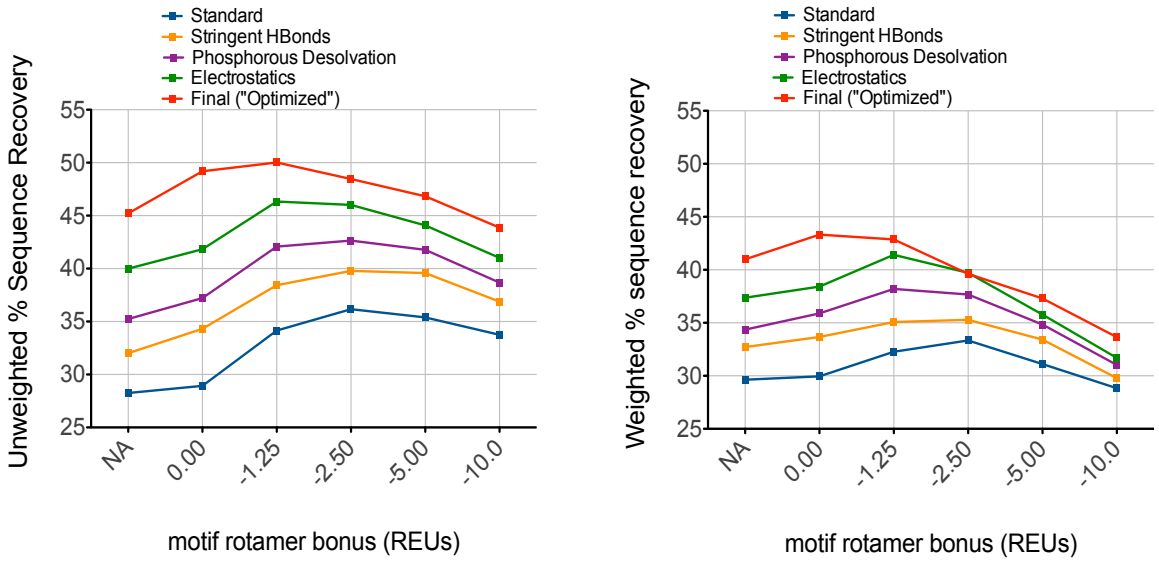


Figure AI.4. Sequence recovery metrics calculated for training set. The full set of 112 proteins was divided into a training set of 48 complexes and a test set of 64 complexes. The sequence recovery for different iterations of energy function optimization and varying motif weight is shown here for the training set. The data from the test set is in the main text.

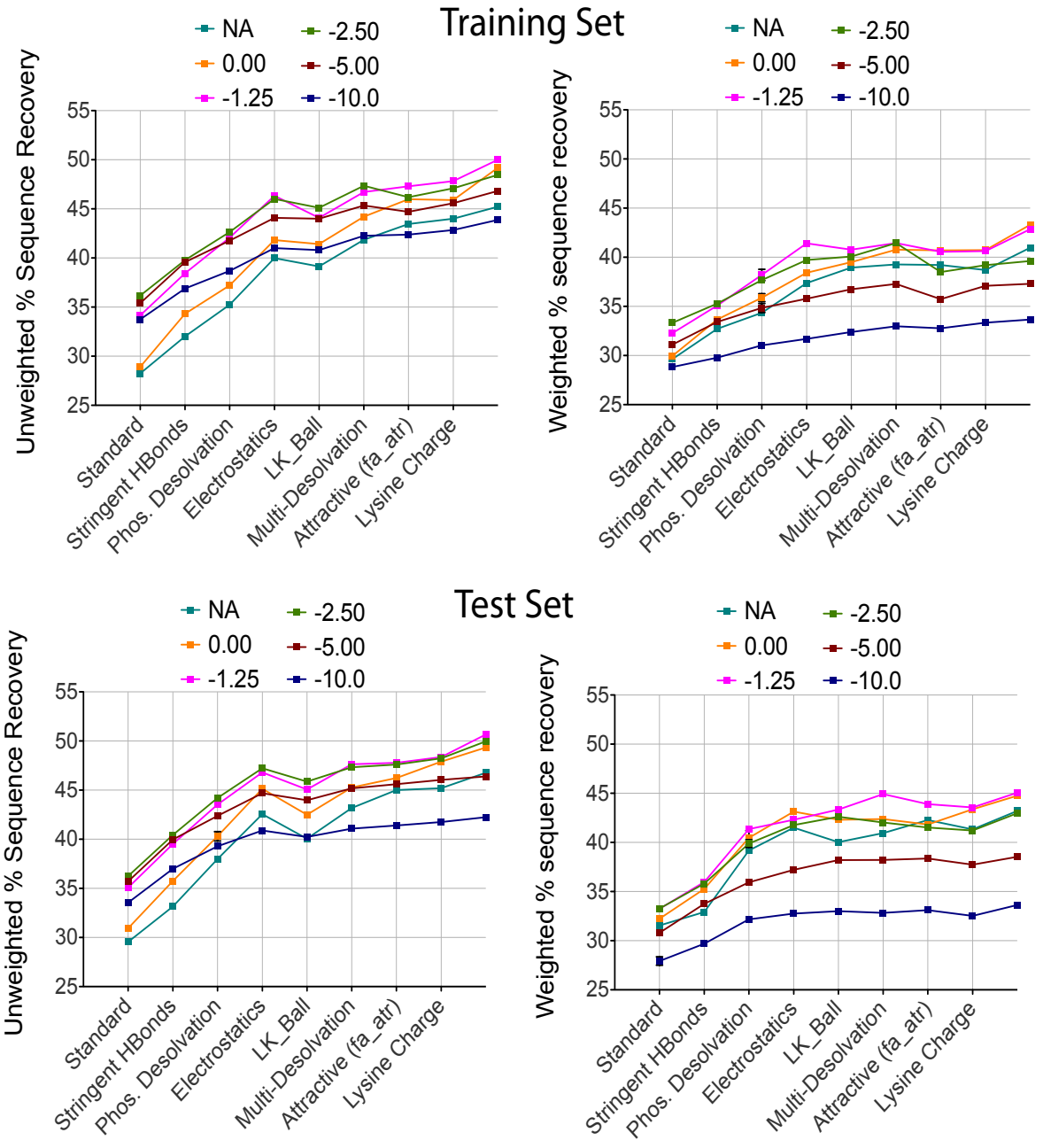


Figure AI.5. Percent recoveries for iterations of ROSETTA energy function optimization. The detailed sequence recovery values for both the test and training sets over all the iterations of energy function optimization and tested motif weights. Each line represents the recovery with a different weight on the added motif rotamers, and NA is the recovery with no motif rotamers added.

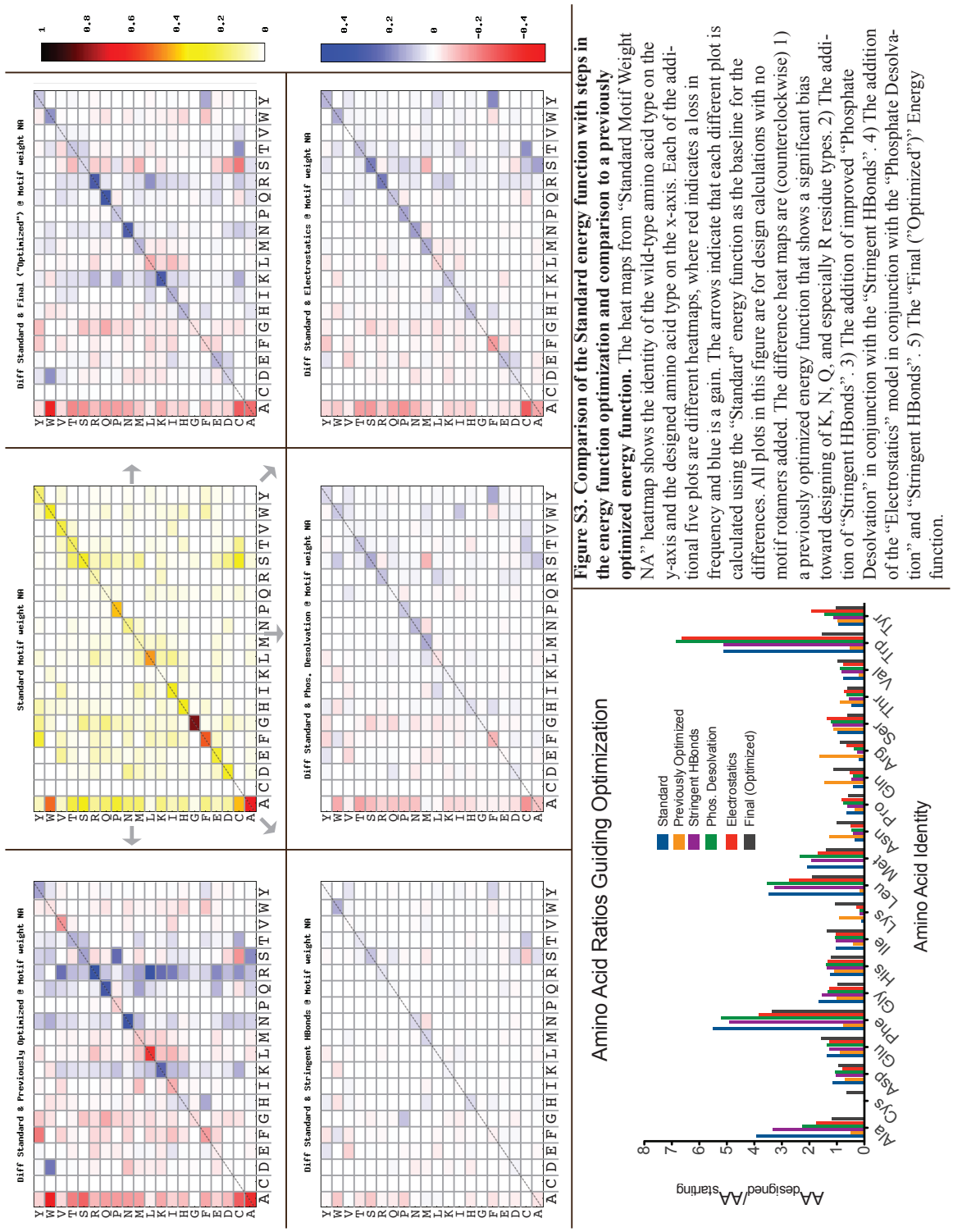


Figure S3. Comparison of the Standard energy function with steps in the energy function optimization and comparison to a previously optimized energy function. The heat maps from "Standard Motif Weight NA" heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the "Standard" energy function as the baseline for the differences. All plots in this figure are for design calculations with no motif rotamers added. The difference heat maps are (counterclockwise) 1) a previously optimized energy function that shows a significant bias toward designing of K, N, Q, and especially R residue types. 2) The addition of "Stringent HBonds". 3) The addition of improved "Phosphate Desolvation" in conjunction with the "Stringent HBonds". 4) The addition of the "Electrostatics" model in conjunction with the "Phosphate Desolvation" and "Stringent HBonds". 5) The "Final ("Optimized")" Energy function.

Figure AI.6

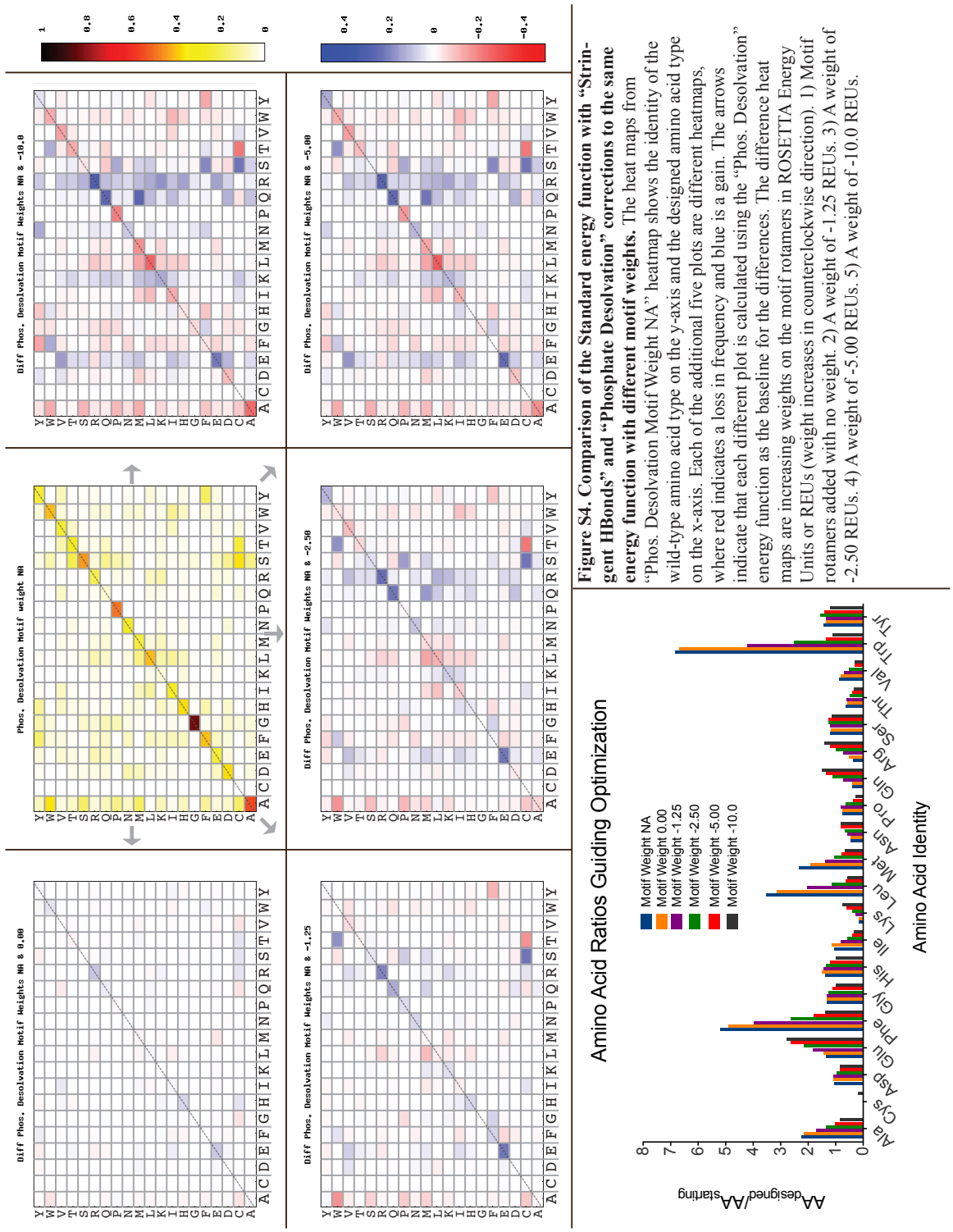


Figure AI.7

Figure S4. Comparison of the Standard energy function with "Stringent HBonds" and "Phosphate Desolvation" corrections to the same energy function with different motif weights. The heat maps from "Phos. Desolvation Motif Weight NA" heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the "Phos. Desolvation" energy function as the baseline for the differences. The difference heat maps are increasing weights on the motif rotamers in ROSETTA Energy Units or REUs (weight increases in counterclockwise direction). 1) Motif rotamers added with no weight. 2) A weight of -1.25 REUs. 3) A weight of -2.50 REUs. 4) A weight of -5.00 REUs. 5) A weight of -10.0 REUs.

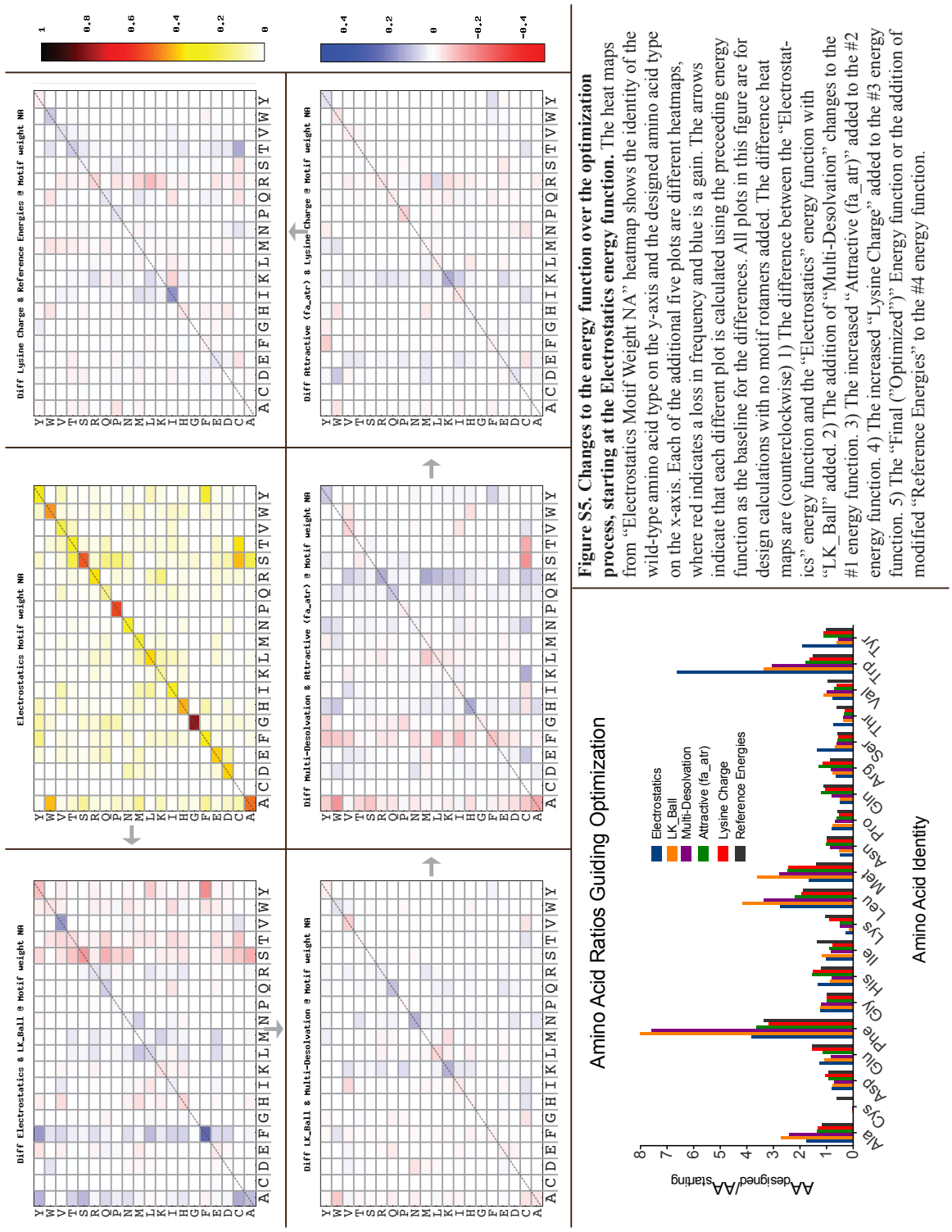


Figure AI.8

Figure S5. Changes to the energy function over the optimization process, starting at the Electrostatics energy function. The heat maps from "Electrostatics Motif Weight NA" heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the preceeding energy function as the baseline for the differences. All plots in this figure are for design calculations with no motif rotamers added. The difference heat maps are (counterclockwise) 1) The difference between the "Electrostatics" energy function and the "Electrostatics" energy function with "LK_Ball" added. 2) The addition of "Multi-Desolvation" changes to the #1 energy function. 3) The increased "Attractive (fa_atr)" added to the #2 energy function. 4) The increased "Lysine Charge" added to the #3 energy function. 5) The "Final ("Optimized")" Energy function or the addition of modified "Reference Energies" to the #4 energy function.

Table AI.4. Summary of I-AniI interface randomization sequencing data.

Pos/AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Tot.
18	2	1														1	2		12	2	20
20	3					13										1	1	3			21
22	2	1				11	1							2		3	1				21
24							12		9												21
25	4		1	1		2				1				1	3	3	2	1			19
26	1					4	1	1	1				2		3	1	4	1	1	1	21
27					1		7			2									4	7	21
29	10					3				2						3	3				21
31	2	2		3				2			2		1				8	1			21
33	4					11										5					20
35	2	1	1	7						1				2		6		1			21
37	5	1		1		6									3	2		3			21
55	1	1		1	1	2		2			2		2		1	1	4	2			20
57	1	2				1					2			1	1	5	4	2		1	20
59									1						20						21
61							1		1						18			1			21
64	1	1			1	1		1		1		1		3	2	4	3	1		1	21
66	3	1								4		1		2	1	1	4	2		2	21
68	2					3			2	1	1				5	2	3	3			22
70										1					20						21
72															21						21
148				19																1	20
150		17														5					22
152	3	1				1		1				1	1		1	4	5			1	19
154		2	1				3									1				14	21
156	2					2			1	1			2	1	3		7	1			20
157	3				1	1	1			3	1	1			5	3	2				21
160	3	1	1			1	2			2		1	2		1	2	1	2	2		21
162	1	1	1		7													1		7	18
164	4	3						1		1	2	1		1		1	3	3			20
166	4		2			1						5				8					20
168		6	8		1	1										5					21
170	6	2				7			1							5					21
172					1	3	1			3		2		1	4	4	1	1			21
189	3			1		4				2			2			4	4	1			21
192	3	3				7	1			2	1				2	1			1		21
194	3	4	2			3	1					1	3			5	1				23
196	4			1		4	5		2	2					5	1	3		1	1	29
198								1		6	1		2				10	6			26
200					1		3		1			2			12			1		2	22
202									20												20
204	1			1		7						1	2			3	1	4			20
205	3	1			1	7	1		1			3	1	1	1	1					21
243															21						21

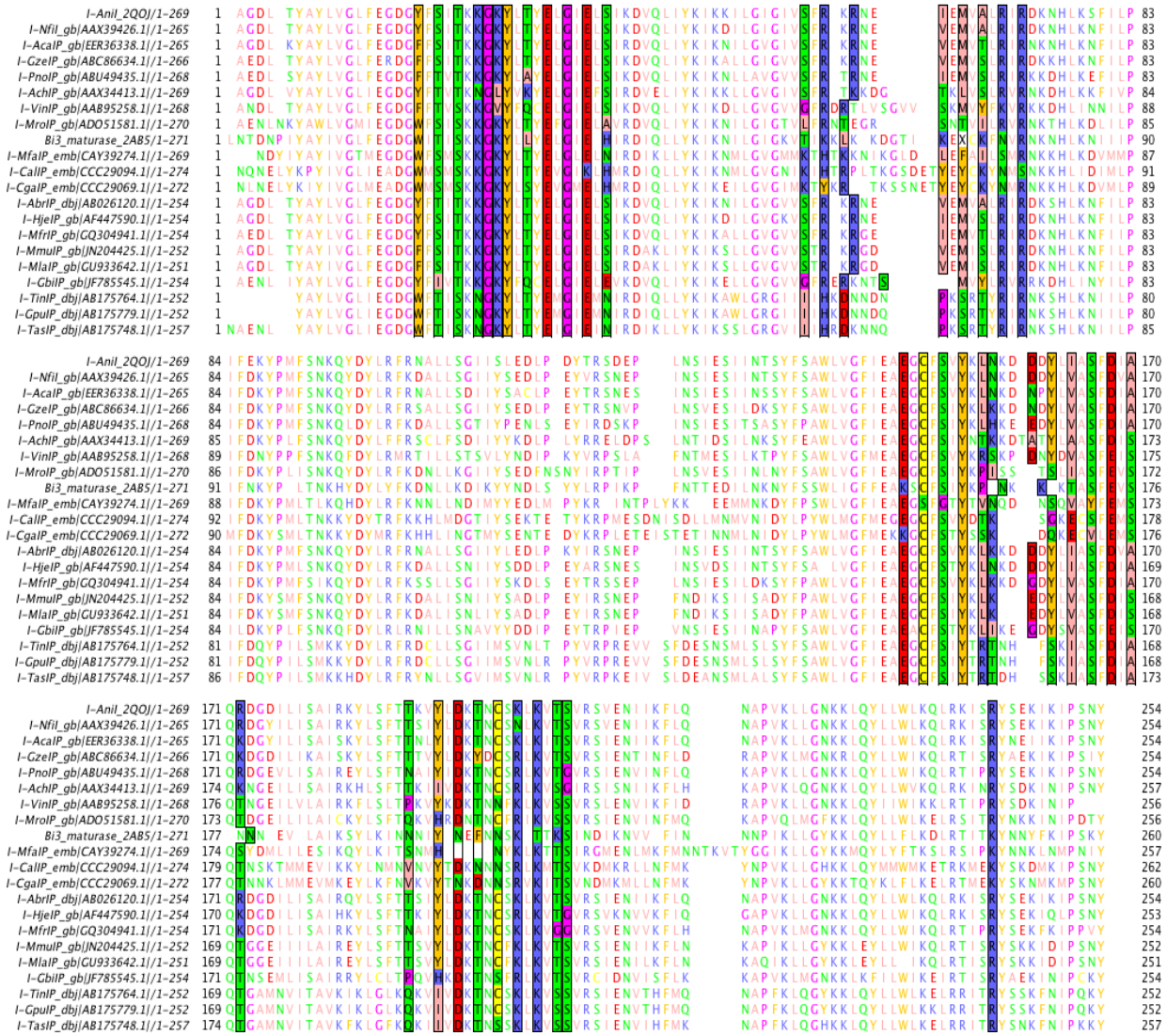


Figure AI.9. Alignment of proteins homologous to I-AniI.
Multiple sequence alignment of proteins homologous to I-AniI found using NCBI blastp and tblastn. The most distant homologue in this alignment, I-MfaIP, has 47% sequence identity and 70% similarity. Alignments with less than 40% identity were predicted to have significantly divergent putative target site sequences (data not shown) and were not chosen for examination. The 44 positions that were randomized in the bacterial selection experiment described in this paper are boxed and highlighted in this multiple sequence alignment. Many of these positions show less variability in the alignment compared to the selection results, most likely due to the differences in conditions and selection pressures between the engineering experiment and natural evolution. It is important to point out as well that, while these enzymes most likely have some activity on the native I-AniI target site, it is probable that enzymes with positions in this alignment that have a high abundance of an amino acid type not observed in the selection experiment are likely targeting substrates with single or multiple nucleotide substitutions.

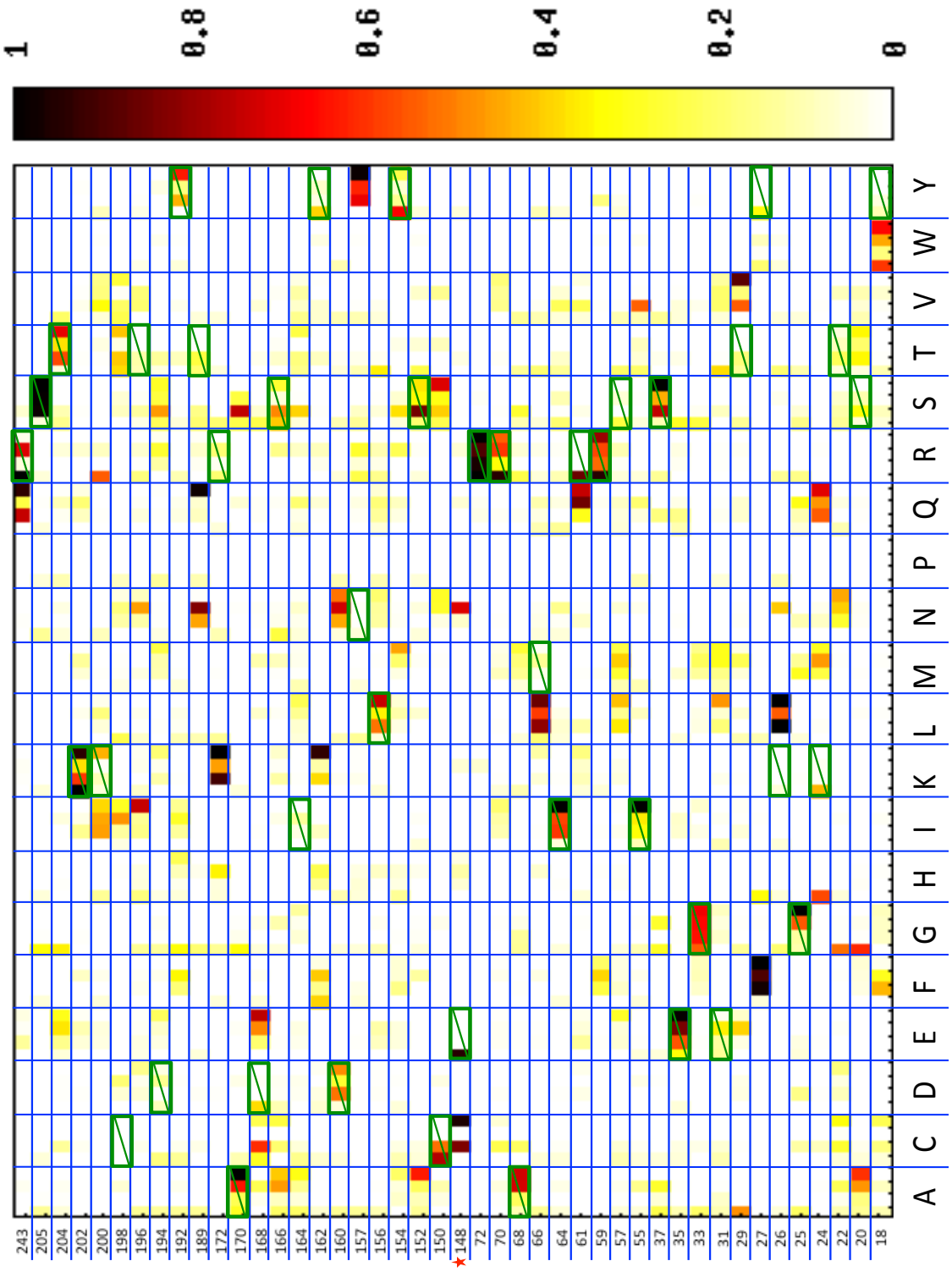


Figure AI.10. Heatmap comparison of experimental data with predictions from three computational methods. Frequencies of amino acid occurrence for each I-AniI interface position in 1) Experimental data 2) HighTemp-Packer 3) DNA-Rebuild 4) Standard with "Optimized" energy function. Wild-type is boxed in green and the order in the boxes is 1-4 from left to right. Each computational protocol was run 56 times.

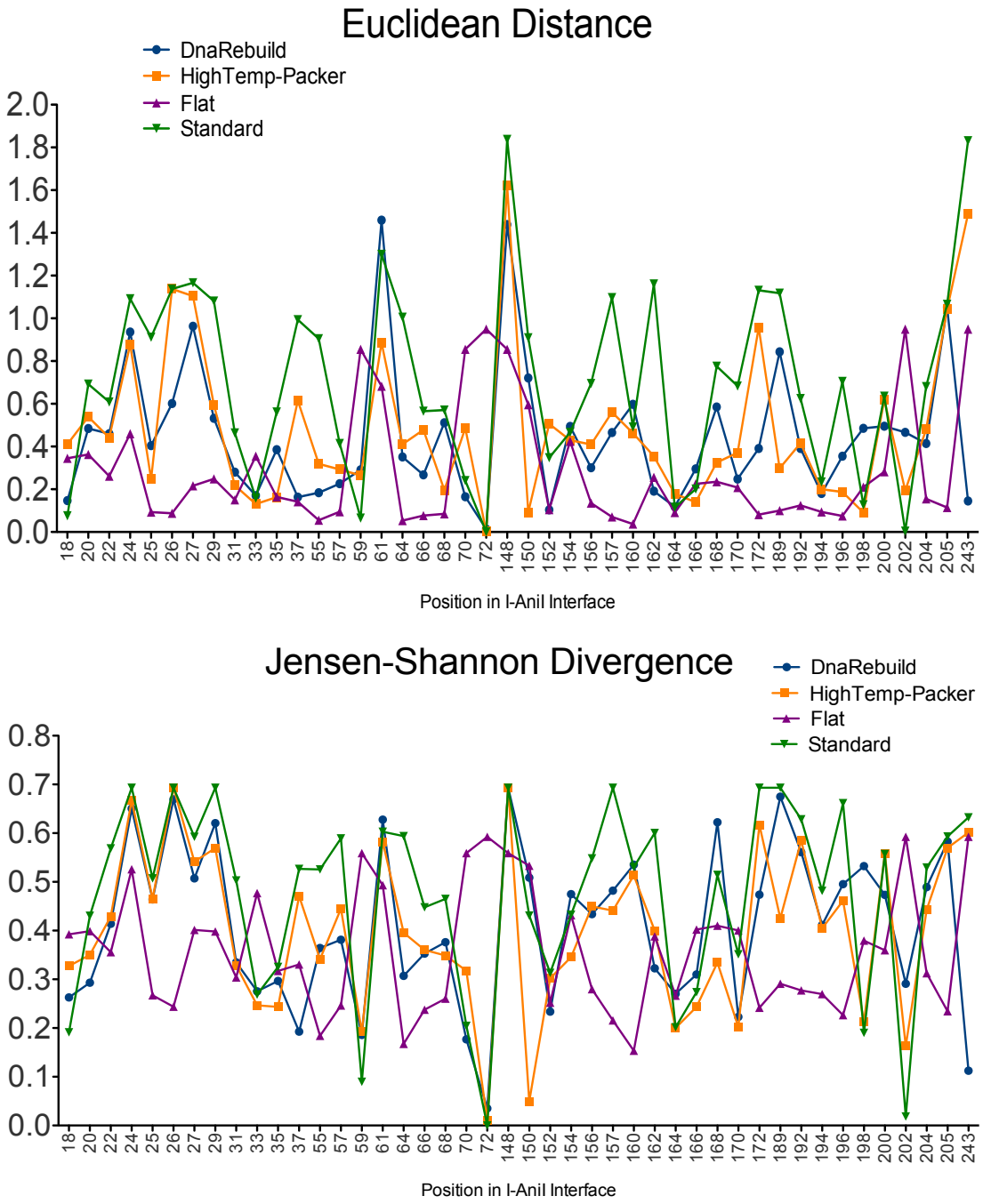


Figure AI.11. Comparison of experimental data with predictions from three computational methods using Euclidean distance and Jensen-Shannon divergence. Positions in the I-AniI interface that were randomized and screened for sequence tolerances are shown on the X-axis. The divergence values for comparisons between the predicted distributions and the experimentally derived distributions were calculated for each randomized position. Lower values of divergence indicate a better match between the calculated and experimental distribution. The "Flat" distribution is a distribution with 0.05 at every position, showing that the computational prediction at many positions is still not displaying as high of sequence diversity as the experimental data.

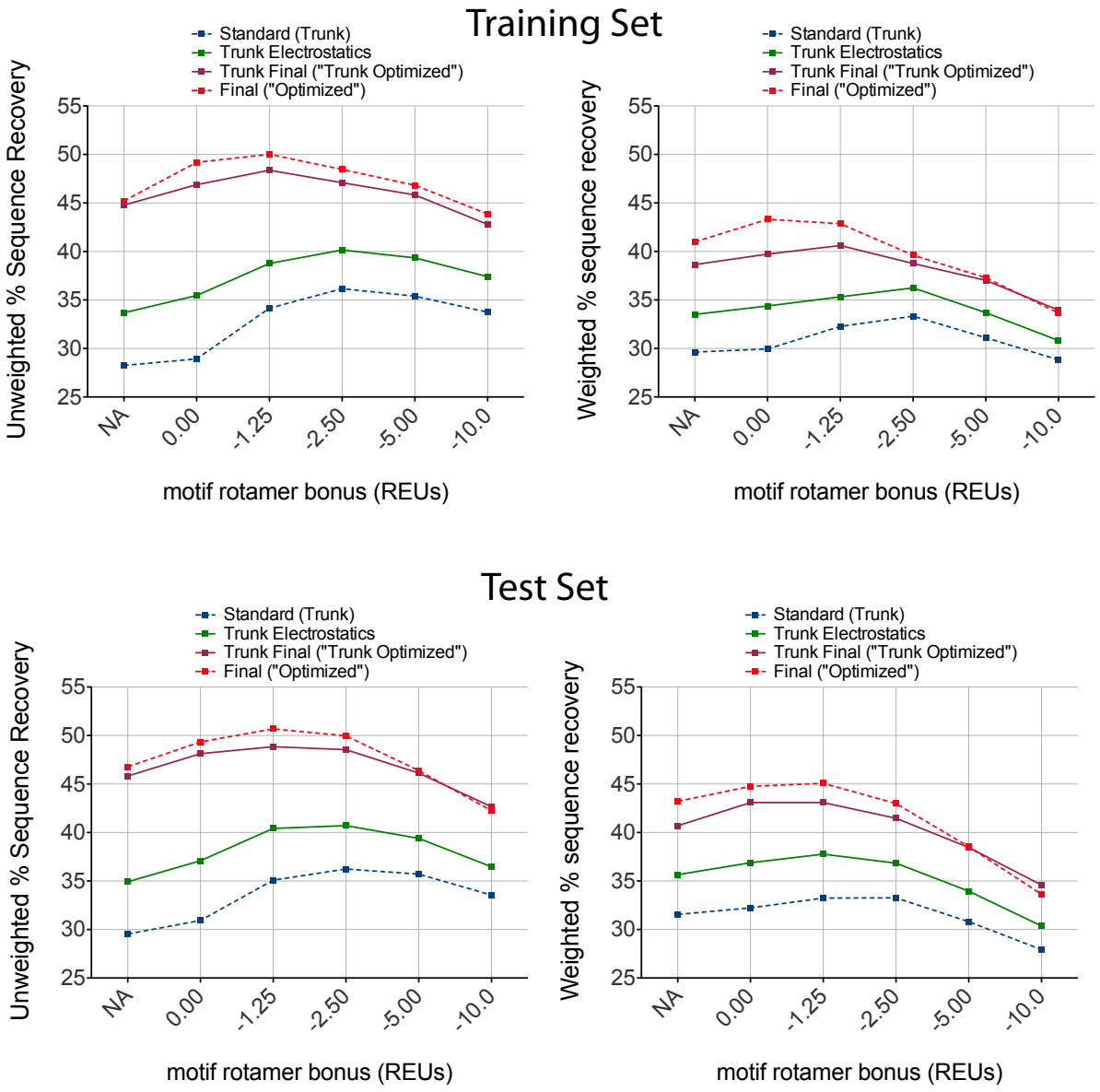


Figure AI.12. Corresponding energy function optimizations in the Trunk version of ROSETTA. The optimizations to the ROSETTA energy function discussed in this work were completed in a branch of ROSETTA with a focus on protein-DNA interactions. The ROSETTA energy function was additionally optimized in the context of the "Trunk" or main version of ROSETTA. The stringent hydrogen bonds and orientation-dependent desolvation model were not available in the "Trunk" version when these calculations were completed. The differences between the "Trunk Optimized" and "Final ("Optimized")" energy functions are the orientation-dependent desolvation, the stringent hydrogen bonds, and differences in the amino acid specific references energies to account for the missing terms. The dashed lines indicate energy functions discussed in the main text of the paper, included for comparison to the trunk optimizations.

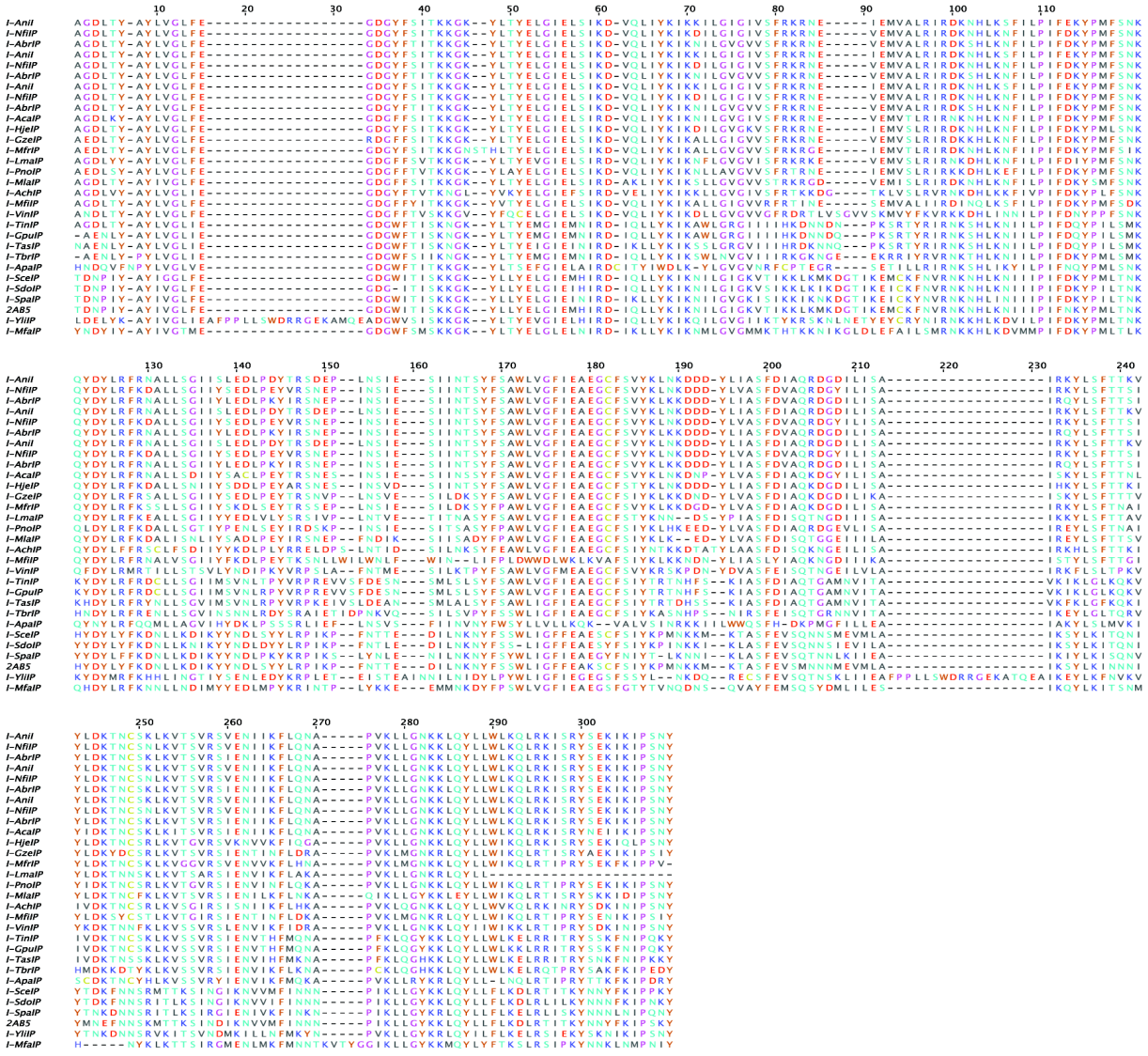


Figure AI.13. Alignment of proteins homologous to I-AniI.
Multiple sequence alignment of proteins homologous to I-AniI found using NCBI BLAST's blastp and tblastn ([g]) that have putative (denoted by suffix P ([h])) endonuclease activity as of March 2011. The repository of data from new protein sequences and their characterizations is continuously growing (e.g. the recently elucidated maturase activity of the I-AniI homologue 2AB5 ([i]). The most distant homologue in this alignment, I-MfaIP, has 47% sequence identity and 70% similarity. Alignments with less than 40% identity were predicted to have significantly divergent putative target site sequences (data not shown) and were not chosen for examination.

---

g. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403-410
h. Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.Kh., Dryden, D.T., Dybvig, K., et al. (2003) *Nucleic Acids Res.* **31**, 1805-1812
i. Longo, A., Leonard, C.W., Bassi, G.S., Berndt, D., Krahn J., Hall, T.M., and Weeks, K.M. (2005) *Nat. Struct. Mol. Biol.* **12**, 779-787

Figure AI.14. Example putative target site identification for I-PnoIP, an I-AniI homologue. An intron (approximately spanning positions 8073 through 9092) within the cytochrome B gene of the complete annotated Phaeosphaeria nodorum SN15 mitochondrion genome (gb|EU053989.1) is observed to encode a LAGLIDADG endonuclease from position 8108 through 9037 (bold text), the homologue ORF. The flanking nucleotide sequences were examined and aligned as shown to each half of the native I-AniI target site. The resulting predicted sequence of the target site for I-PnoIP is colored yellow.

Table AI.5. Percent identity and similarity to I-AniI for each homologue previously identified and aligned in Figure AI.13.

Homologue Name	Accession Number	Percent Identity to I-AniI	Percent Similarity to I-AniI
I-AniI	pdb|2QOJ|Z	100%	100%
I-NfiIP	gb|AAX39426.1	93%	98%
I-AbrIP	dbj|AB026120.1	92%	97%
I-AcaIP	gb|EER36338.1	88%	95%
I-HjeIP	gb|AF447590.1	86%	95%
I-GzeIP	gb|ABC86634.1	84%	93%
I-MfrIP	gb|GU952815.1	81%	90%
I-LmaIP	emb|CBX89979.1	79%	89%
I-PnoIP	gb|ABU49435.1	78%	91%
I-MlaIP	gb|GU933642.1	72%	86%
I-AchIP	gb|AAX34413.1	71%	86%
I-MfiIP	gb|AF343070.1	69%	79%
I-VinIP	gb|AAB95258.1	64%	79%
I-TinIP	dbj|AB175764.1	59%	75%
I-GpuIP	dbj|AB175779.1	59%	75%
I-TasIP	dbj|AB175746.1	57%	74%
I-TbrIP	dbj|AB175750.1	57%	76%
I-ApaIP	gb|AF538047.1	52%	67%
I-SceIP	emb|CAA32785.1	51%	70%
I-SdoIP	emb|X59280.1	50%	72%
I-SpaIP	gb|EU852811.1	50%	72%
2AB5 (maturase)	pdb|2AB5|A	49%	67%
I-YliIP	emb|AJ307410.1	49%	66%
I-MfaIP	emb|CAY39274.1	47%	70%

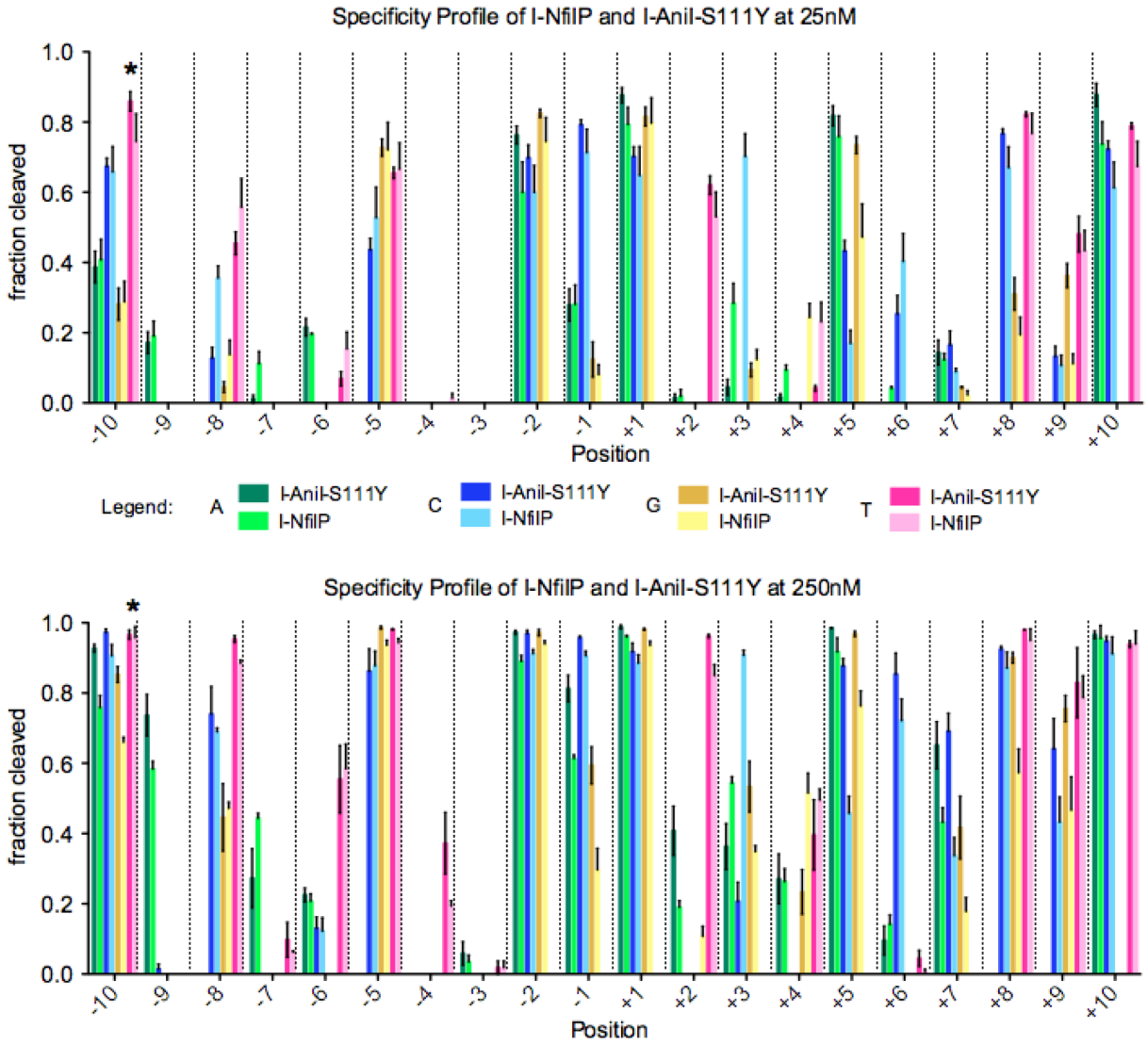


Figure AI.15. Specificity profile of I-NfiIP and I-AniI-S111Y at 25nM and 250nM enzyme concentration. Cleavage fractions of wild-type (denoted by asterisk) and singly-substituted target sites for all 20 positions are shown. A fraction of 1 indicates complete cleavage of the substrate with no remaining uncut fraction. Since the sequence of homologue I-NfiIP contains the activating mutation S111Y, its profile is compared to that of I-AniI-S111Y. Though the specificity pattern is similar overall, the observed novel +3C specificity was later confirmed in a variant containing the I-NfiIP-derived K200N transfer (Figure 21b).

Table AI.6. $EC_{0.5max}$ (nM) cleavage efficiencies and cleavage plateaus (f_{max} , the maximal fraction of site cleavage) for homologue-based I-AniI variants tested on singly-substituted target sites. Variants are grouped into categories (C-terminal loops, K200, Central 4 loops, Core Mutations) dependent on the location and theorized role of the mutations transferred to the I-AniI scaffold. All variants include the F80K and L233K mutations for solubility, with the exception of the K24N/T29K variant that only includes F80K. The base activity column indicates whether the variant was made with the activating F13Y or S111Y mutations. $EC_{0.5max}$ values are the mean (nM) ± coefficient of variation (%CV) of two independently determined enzyme cleavage profiles chosen (via inspection for substrate degradation and experimental error) from at least two separate *in vitro* cleavage assays on plasmid DNA substrates containing single base-pair substitutions from the I-AniI wild-type target site. The CV is given as a percentage measure of variability estimated through dividing the standard error from the mean (SEM) by the mean. Values of $EC_{0.5max}$ >750nM are too high to allow accurate quantitative determination, and therefore no cleavage plateau is reported. $f_{max} \approx 1$ designates a cleavage plateau at its greatest allowable value, meaning the substrate can be completely cleaved with no remaining uncut fraction.

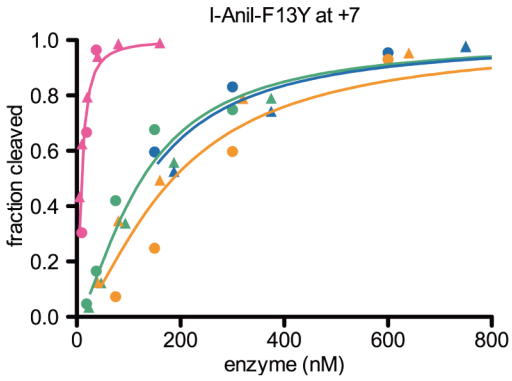
Variant	Mutations	Base Activity	Position		$EC_{0.5max}$ (nM)	Cleavage Plateau
C-terminal loops						
I-PnoIP	V153I, N157H, D159E, D160E, I164V	F13Y	wild-type		4 ± 3%	0.980
			+8A	C	3 ± 28%	0.987
				G	19 ± 40%	0.938
				T	3 ± 41%	0.975
			+9A	C	70 ± 23%	~1
				G	41 ± 23%	~1
				T	15 ± 31%	0.987
			+10G	A	6 ± 9%	0.981
				C	8 ± 27%	0.960
				T	5 ± 42%	0.974
I-AchIP	V153I, K155N, L156T, N157K, D160T, inserted A after 160, D161T, I164A	F13Y	wild-type		8 ± 6%	0.977
			+7T	A	8 ± 31%	0.995
				C	74 ± 5%	0.965
				G	328 ± 7%	~1
			+8A	C	8 ± 4%	0.985
				G	13 ± 6%	0.948
				T	6 ± 27%	0.980
			+9A	C	10 ± 8%	0.981
				G	11 ± 8%	0.986
				T	9 ± 3%	0.981
			+10G	A	8 ± 20%	0.966
				C	8 ± 13%	0.957
				T	8 ± 23%	0.974
I-TasIP	K155T, L156R, N157T, K158 deletion, D160H, D161S, Y162S, L163K	F13Y	wild-type		86 ± 35%	0.941
			+7T	A	20 ± 32%	0.934
				C	588 ± 26%	~1
				G	459 ± 41%	0.752
			+8A	C	48 ± 18%	0.856
				G	6 ± 13%	0.942
				T	35 ± 13%	0.910

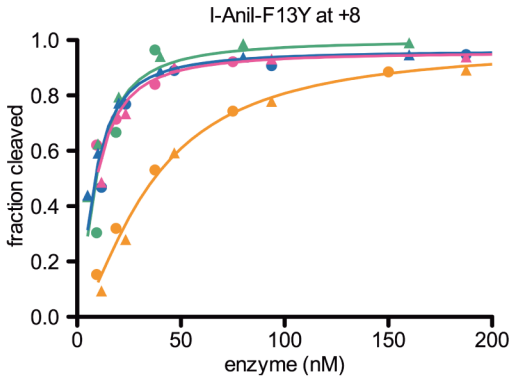
			+9A	C	6 ± 1%	0.857
				G	81 ± 13%	0.864
				T	79 ± 11%	0.925
			+10G	A	139 ± 25%	~1
				C	15 ± 22%	0.928
				T	80 ± 5%	~1
K200						
I-NfiIP	I164V, K200N	F13Y	wild-type		6 ± 5%	0.981
			+3T	A	146 ± 4%	0.854
				C	14 ± 39%	0.925
				G	155 ± 8%	0.670
			+4C	A	260 ± 7%	0.565
				G	205 ± 14%	~1
				T	155 ± 32%	0.612
I-PnoIP	K200R	F13Y	wild-type		7 ± 15%	0.977
			+3T	A	140 ± 7%	~1
				C	358 ± 27%	0.929
				G	42 ± 9%	0.996
			+4C	A	118 ± 4%	~1
				G	116 ± 8%	~1
				T	18 ± 17%	0.962
			+5T	A	5 ± 28%	0.951
				C	17 ± 31%	0.950
				G	9 ± 24%	0.933
Central 4 loops						
I-GzeIP*	R172K, T196Y, N197D	S111Y	wild-type		7 ± 28%	0.919
			+2C	A	100 ± 7%	~1
				G	348 ± 7%	0.717
				T	27 ± 25%	0.954
			+3T	A	212 ± 6%	~1
				C	315 ± 6%	0.860
				G	156 ± 6%	~1
I-AchIP	K60T, R61K, N62K, E63D, inserted G after 63, I64T, E65K, M66L	S111Y	wild-type		18 ± 4%	0.956
			-2T	A	6 ± 2%	0.953
				C	40 ± 6%	0.987
				G	10 ± 16%	0.979
I-VinIP Loop	S57G, K60D, N62T , E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	wild-type		32 ± 9%	0.962
			-2T	A	26 ± 5%	0.927
				C	21 ± 24%	0.929
				G	23 ± 28%	0.932
Core mutations						
I-VinIP w/o core	A68Y, R70K		wild-type		>750	N/A
			-6G	A	>750	N/A
				C	>750	N/A
				T	>750	N/A
			-5A	C	>750	N/A
				G	>750	N/A
				T	>750	N/A
I-VinIP w/ core	A68Y, L69F, R70K, I71V		wild-type		428 ± 2%	~1
			-6G	A	>750	N/A
				C	>750	N/A

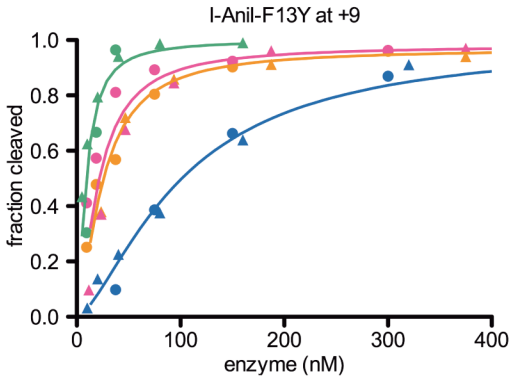
				T	>750	N/A
			-5A	C	>750	N/A
				G	721 ± 5%	~1
				T	>750	N/A
I-VinIP-S111Y w/o core	A68Y, R70K	S111Y	wild-type		68 ± 16%	~1
			-6G	A	182 ± 45%	0.229
				C	>750	N/A
				T	245 ± 15%	~1
			-5A	C	523 ± 3%	~1
				G	111 ± 23%	~1
				T	424 ± 4%	0.955
I-VinIP-S111Y w/ core	A68Y, L69F, R70K, I71V	S111Y	wild-type		31 ± 34%	~1
			-6G	A	176 ± 24%	0.310
				C	>750	N/A
				T	135 ± 13%	0.983
			-5A	C	362 ± 0%	~1
				G	89 ± 10%	0.996
				T	404 ± 1%	~1
I-VinIP Loop	S57G, K60D, N62T , E63L, I64V, SGVVS insert after 64, E65K, A68Y, L69F, R70K, I71V	S111Y	wild-type		32 ± 9%	0.962
			-6G	A	208 ± 22%	0.995
				C	112 ± 19%	0.996
				T	28 ± 13%	0.994
			-5A	C	67 ± 1%	0.989
				G	31 ± 8%	0.957
				T	38 ± 15%	0.969
K24N/L28V/T29K	K24N, L28V, T29K		wild-type		>750	N/A
			-8A	C	>750	N/A
				G	23 ± 15%	0.939
				T	750 ± 9%	~1
K24N/T29K	K24N, T29K, lacking L233K		wild-type		>750	N/A
			-8A	C	>750	N/A
				G	23 ± 8%	0.873
				T	>750	N/A
I-AniI Base Activity Comparison						
I-AniI-F13Y	F13Y	F13Y	wild-type		10 ± 31%	0.996
			+2C	A	194 ± 35%	0.982
				G	608 ± 11%	0.310
				T	17 ± 23%	0.900
			+3T	A	343 ± 26%	0.950
				C	503 ± 20%	0.869
				G	214 ± 7%	~1
			+4C	A	393 ± 13%	0.921
				G	437 ± 25%	0.898
				T	236 ± 19%	0.943
			+5T	A	11 ± 24%	0.934
				C	60 ± 35%	0.967
				G	27 ± 40%	0.987
			+7T	A	119 ± 28%	0.949
				C	140 ± 21%	~1
				G	189 ± 26%	~1
			+8A	C	9 ± 28%	0.964

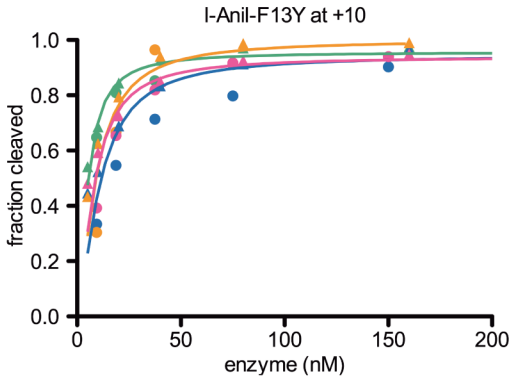
				G	36 ± 11%	0.983
				T	9 ± 26%	0.956
			+9A	C	101 ± 1%	~1
				G	25 ± 10%	0.971
				T	23 ± 44%	0.981
			+10G	A	5 ± 9%	0.959
				C	12 ± 35%	0.950
				T	9 ± 36%	0.944
I-AniI-S111Y	S111Y	S111Y	wild type		9 ± 21%	0.969
			-6G	A	>750	N/A
				C	379 ± 21%	~1
				T	173 ± 5%	~1
			-5A	C	55 ± 2%	~1
				G	21 ± 21%	0.989
				T	27 ± 25%	0.985
			-2T	A	7 ± 20%	0.959
				C	12 ± 1%	0.967
				G	11 ± 3%	0.975

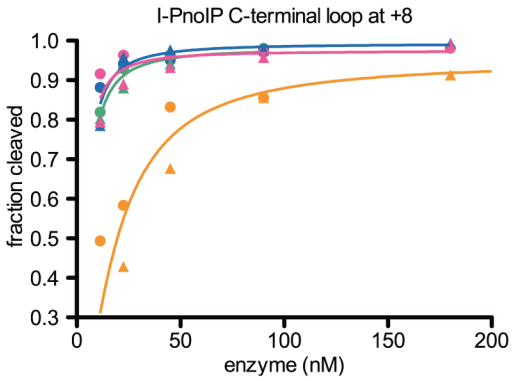
* This variant contains an additional I248V mutation that does not affect cleavage efficiency, as it is at the end of the C-terminal domain far from the interface.

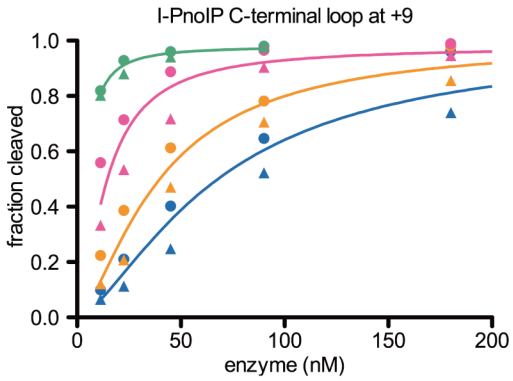

I-AniI-F13Y at +7
fraction cleaved
enzyme (nM)

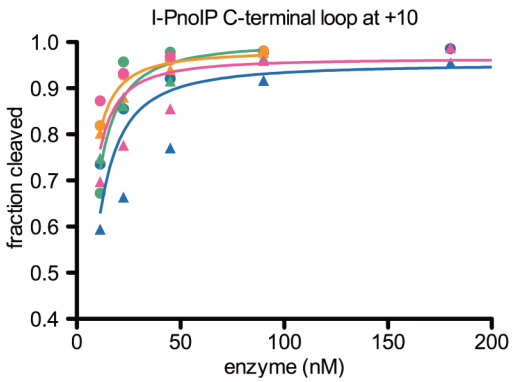

I-AniI-F13Y at +8
fraction cleaved
enzyme (nM)


I-AniI-F13Y at +9
fraction cleaved
enzyme (nM)


I-AniI-F13Y at +10
fraction cleaved
enzyme (nM)


I-PnoIP C-terminal loop at +8
fraction cleaved
enzyme (nM)


I-PnoIP C-terminal loop at +9
fraction cleaved
enzyme (nM)


I-PnoIP C-terminal loop at +10
fraction cleaved
enzyme (nM)

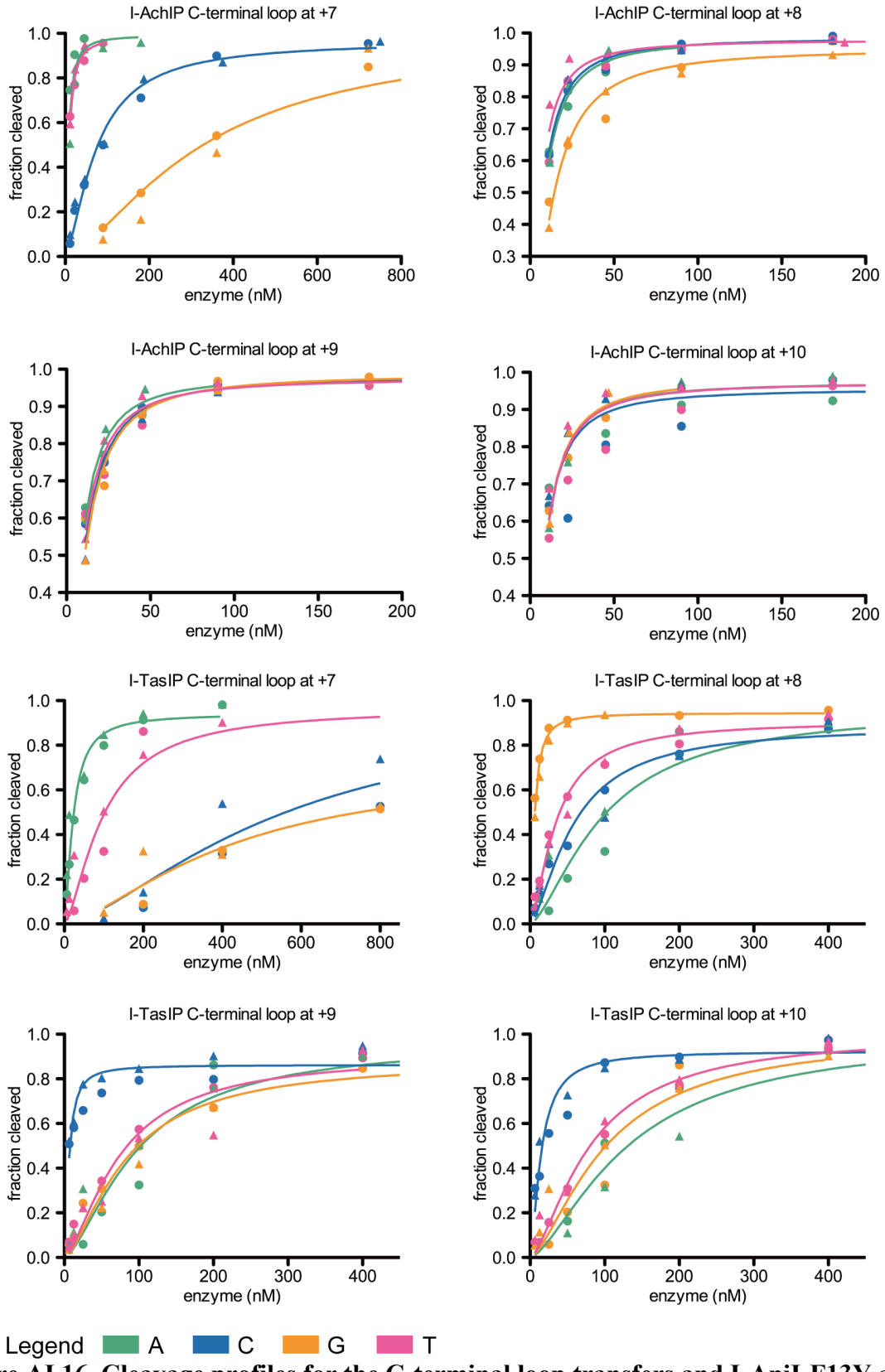


Figure AI.16. Cleavage profiles for the C-terminal loop transfers and I-AniI-F13Y at positions +7 through +10.

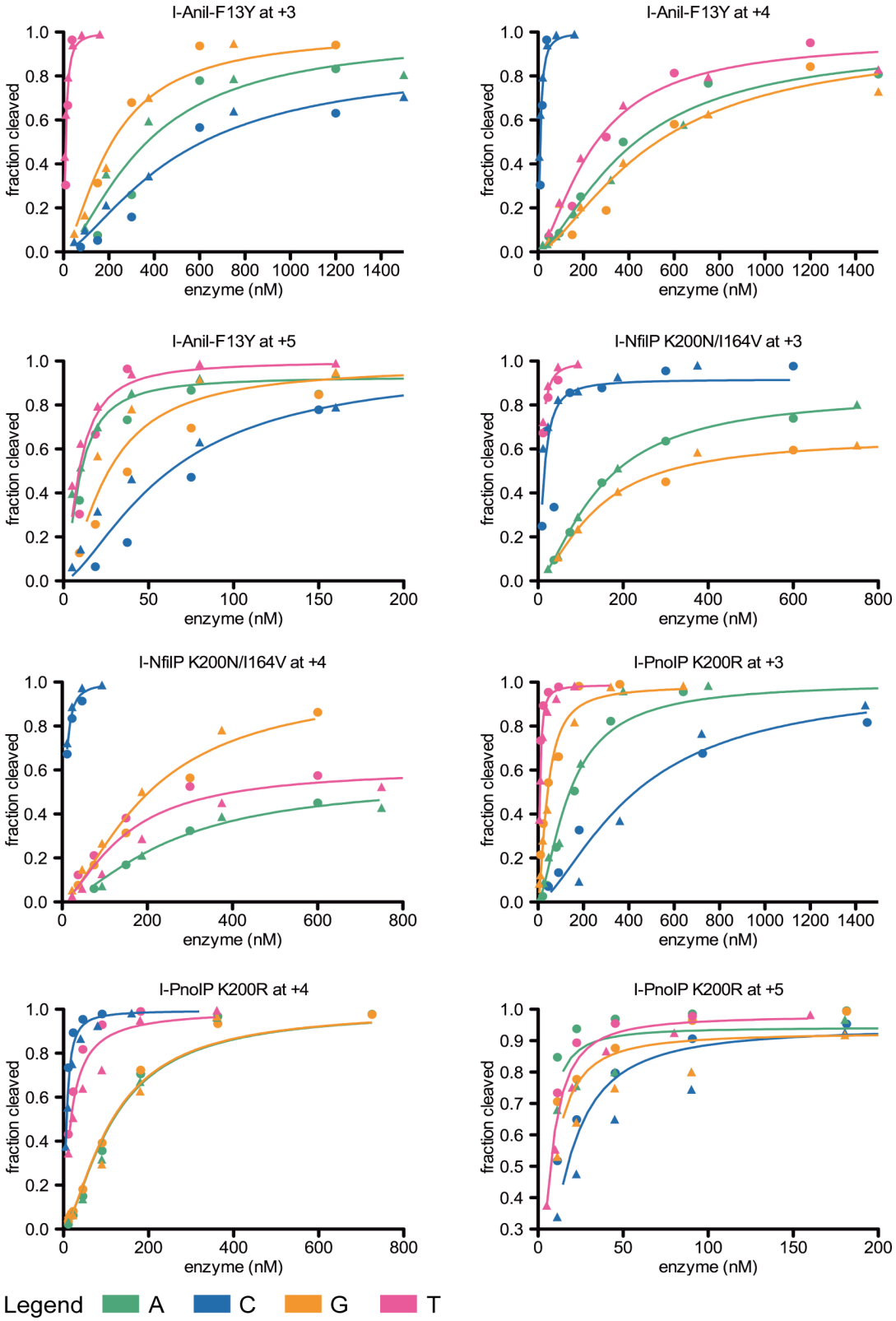


Figure AI.17. Cleavage profiles for the K200 variants and I-AniI-F13Y at +3, +4, and +5.

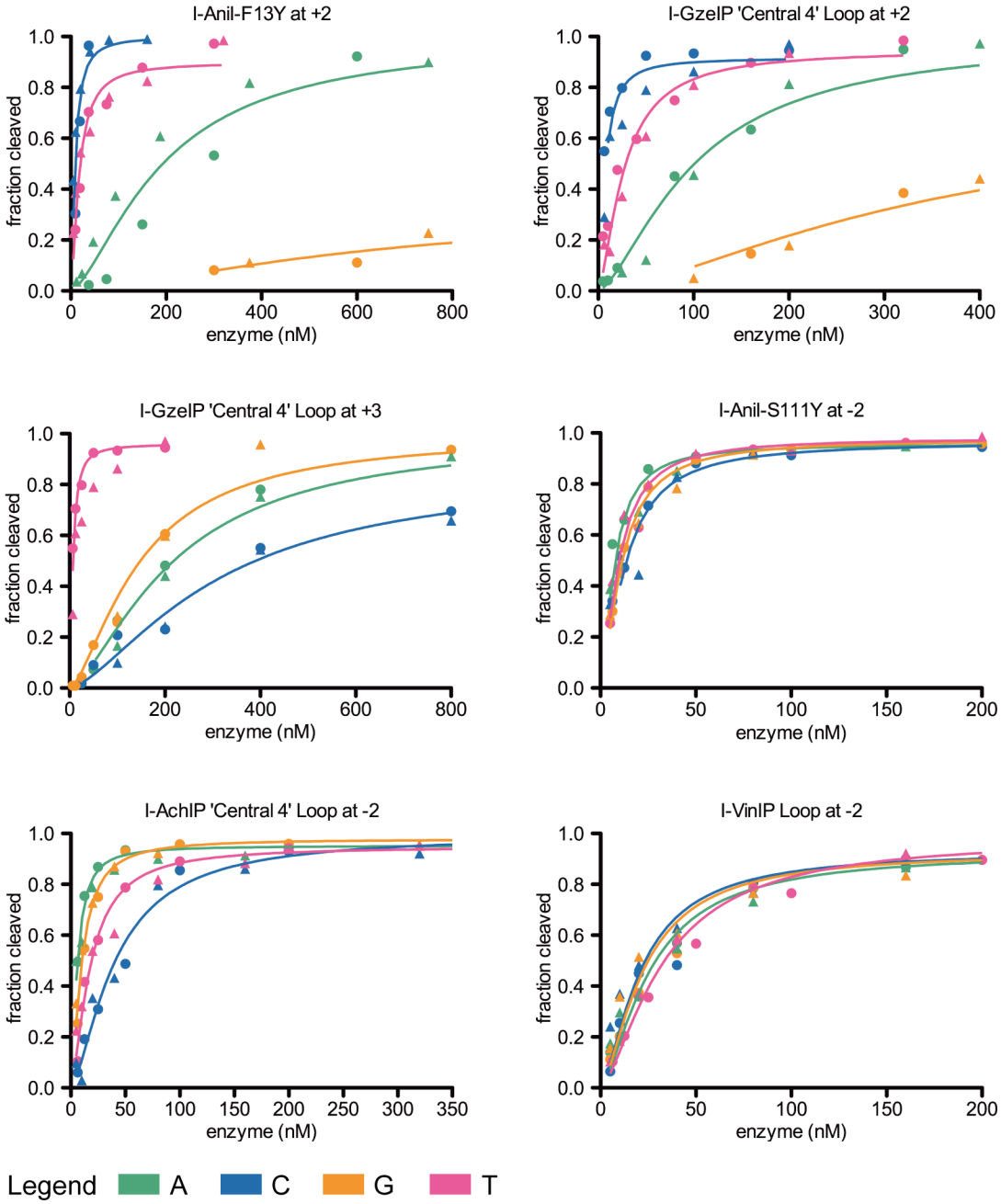


Figure AI.18. Cleavage profiles of variants affecting the central four target site positions. Results for I-AniI-F13Y at position +2 and I-AniI-S111Y at -2 are provided here for comparison; the plot for I-AniI-F13Y at +3 can be seen in Figure AI.16.

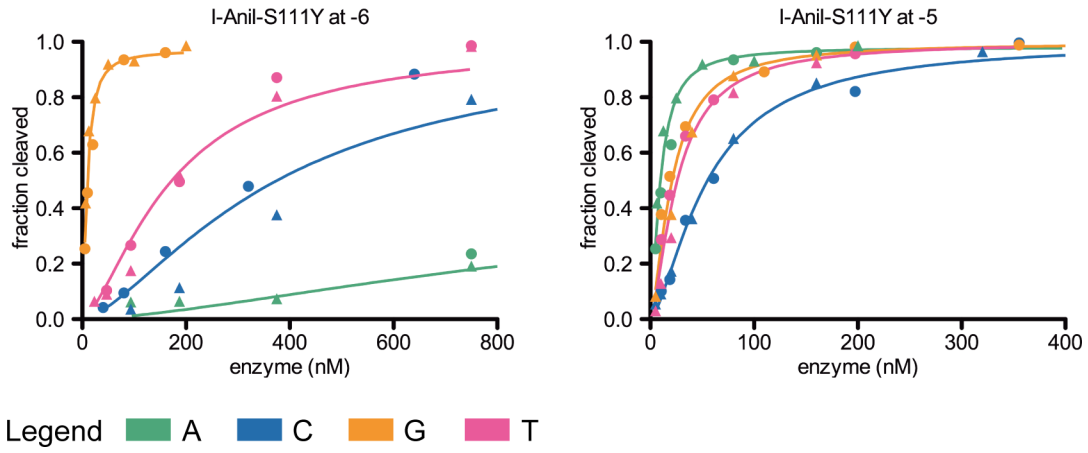


Figure AI.19. Cleavage profiles for I-AniI-S111Y at positions -6 and -5. The corresponding $EC_{0.5max}$ (nM) were shown previously (Figure 22b) along with the cleavage plots for the I-VinIP derived variants (Figure 22a) evidencing the important role of core mutations.

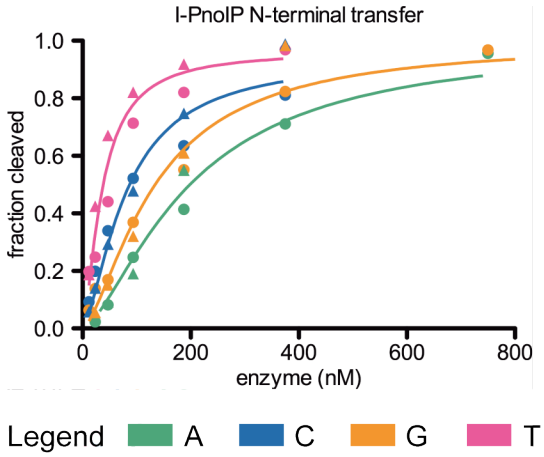


Figure AI.20. Cleavage profile for I-PnoIP N-terminal transfer at the -5 position. The -5T substitution is preferred over the wild-type -5A. This result is in accordance with the -5T of the predicted I-PnoIP target site given in Figure 20c.

Appendix II: Detailed Methods

Appendix II.1 Methods for Chapter 2

Protein expression and purification

Genes encoding I-AniI [1] designs were assembled from oligonucleotides [2], cloned into a variant of the pet15 expression vector, and sequence-verified plasmids were transformed into BL21 Star (Invitrogen). A one litre culture of auto-induction media [3] was inoculated with several colonies, grown at 37°C for ca. 12 hours (to approximately saturation), and expression at 18°C was continued for ca. 24 hours. Cells were harvested, resuspended in Tris 20 mM pH 7.5, 1.0 M NaCl, and 30 mM Imidazole, lysed by sonication and lysozyme. The soluble fraction was loaded onto a 1 mL HisTrap FF crude column (GE Healthcare) and I-AniI variants were purified by Imidazole gradient elution on an AKTA express (GE Healthcare). The proteins were concentrated and the buffer was exchanged to Tris 20 mM pH 7.5, 500 mM NaCl, and 50% (v/v) glycerol for storage. Purity of the proteins was assessed by SDS-PAGE gel and the concentration of samples with ca. > 95% purity was determined by measuring the absorbance at 280 nm using the calculated extinction coefficient [4]. The concentration of enzyme in the < 95% pure samples was determined by generating a standard curve of with a pure I-AniI protein, correlating protein concentration with band density (calculated with ImageJ [5]), and comparing the band density of the I-AniI protein in impure samples run on the same gel as the standard curve.

Plasmid substrate construction

All single base-pair variants from the wild-type target site in pBluescript were individually constructed by site-directed mutagenesis as described [6]. Sequence-verified plasmids were linearized with ScaI prior to the kinetic assays to facilitate product identification.

Endonuclease activity assays

Kinetic assays

Previous work [7] confirms that I-AniI, similar to other LAGLIDADG endonucleases, is a single-turnover enzyme, and the conditions for single-turnover kinetics [8] were met in all experiments. The ionic strength of the enzyme reaction buffer was optimized for enzyme activity and stability to a final solution of 170 mM KCl, 10 mM $MgCl_2$, and 20 mM Tris pH 9.0. Enzyme was diluted in 1.25X reaction buffer to working concentrations, serial two-fold dilutions were made, and both substrate plasmid and diluted enzyme were incubated separately at 37°C for 1 minute. The appropriate amount of plasmid (1/5 of the reaction volume) was added to each reaction for a final 1X reaction buffer and final plasmid concentration of ca. 5 nM (lowest concentration still readily visible on agarose gel). The plasmid (1/5 of reaction volume) was added to the enzyme (4/5 of reaction volume) to minimize heat loss during the transfer (found to add significant noise to the data). Reactions were halted with 200 mM EDTA, 30 % glycerol, and bromophenol blue. DNA fragments were separated on 1.2% agarose TBE gels, which were then stained in a standard ethidium bromide solution and subsequently destained in water for maximum contrast between DNA and background. All data was collected by integrating the density of the substrate (2,959 bp) and product bands (1,801 bp and 1,158 bp) using ImageJ [5]. The percent product formed is equal to the sum of the density of the two product bands divided by the total sum of the densities of the 3 bands. The progress curves fit to single exponentials for all enzyme concentrations and for all target sites except for several substitutions in the central four base-pairs between the cleavage sites on the two DNA strands (Figure AI.1).

Assays for specificity positions adjacent to designed nucleotide

Two-fold serial dilutions of enzyme from 1500 nM to 11 nM were made in 1.25X reaction buffer and the enzyme was reacted with ca. 5 nM substrate (in 1X reaction buffer) for a ½ hour at 37°C. Reactions were halted and data was analyzed as described above in the "kinetic assays" section.

Fluorescence competition binding assay [9]

Unlabeled DNA oligonucleotides with each of the 60 single base-pair substitutions in the I-AniI target site (wild-type I-AniI site, 5′-TGAGGAGGTTTCTCTGTAAG-3′), a negative control sequence (5′-CTCTTCTTGCATATATCTCC-3′), an unlabeled wild-type site oligo, and a wild-type site oligonucleotide labeled with 5′ Cy3, were synthesized with six consecutive "A" flanking on each end (Integrated DNA Technology, 100-nmole scale, salt-free). Complementary oligonucleotides were ordered for all 63 sites and double stranded target DNA was preparing by annealing equal amounts of complementary strands.

His-tagged I-AniI was immobilized by incubating 200μl of 100 nM I-AniI in TBS/BSA buffer (50mM Tris-HCl (pH 7.5), 150mM NaCl, 0.2% BSA) in wells of Nickel-NTA coated HisSorb plates (Qiagen) for 2 hours at room temperature. Unbound protein was removed and the plates were washed four times with TBS/Tween-20 (50mM Tris-HCl (pH 7.5), 150mM NaCl, 0.05%Tween-20). The immobilized I-AniI in the microtiter plate was incubated for ca. four hours with both 100 nM labeled target DNA duplex and 3 μM (30-fold excess) of one unlabeled duplex per well in 200 ml of binding buffer (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.02 mg/ml poly(dI-dC), 10mM $CaCl_2$). The plates were washed four times with TBS (50mM Tris-HCl (pH 7.5), 150mM NaCl), and the fluorescent signal retained in each well was quantified using a SpectraMax M5/M5^e micro-plate reader (Molecular Devices) (excitation: 510nm, emission: 565nm, cutoff: 550 nm). Additional negative control experiments performed in the absence of the enzyme indicated that no significant detectable fluorescent signal was retained after the protocol described above was completed. Relative binding affinities were calculated using the following equation:

$$\text{Relative binding affinity} = [(F(n) - F(x)) \times F(t)] / [(F(n) - F(t)) \times F(x)],$$

where F(x), F(t), and F(n) indicate fluorescent intensities obtained from wells in which the immobilized protein was incubated with the unlabeled singly-substituted target sites, wild-type target site, and negative control sequence, respectively.

Computational design methods

Single-state design (designs -9C, -8G_C, -6C, +5C, and +8C)

The computational design of homing endonuclease-DNA specificity was performed using the Rosetta design software in a manner that is specifically designed to predict new protein sequences that will bind with high affinity to novel DNA sequences [10]. The prediction of designed proteins with novel interactions to substituted base-pairs in the I-AniI recognition sequence was performed by mutation and Monte-Carlo repacking of amino acid sidechains as described in Ashworth *et al.* 2006 [11]. The template for the design calculations was the crystal structure of the I-AniI-DNA complex (pdb code 2QOJ). Additionally, minor shifts of the protein backbone were modeled only in the vicinity of the designed region using a loop-rebuilding algorithm [12,13]. The specificity of each hypothetical new protein sequence for the intended new DNA recognition sequence was calculated as the Boltzmann probability of the intended complex versus a partition function consisting of each base-pair possibility at the redesigned DNA base-pair [14]. Following design, predicted protein sequences with the most favorable binding energy and highest predicted specificity were reverted position by position to the wild-type amino acid sequence to identify (and revert) designed mutations that did not significantly contribute to the energy or specificity of the designed complex.

Multi-state design (designs -8G_A and -8G_B)

Two base pair positions in the structure were computationally mutated to generate a partial match to a recognition site in the IL-2Rγ gene in a mouse model of severe combined immunodeficiency disease (SCID). Specifically, positions -9G:C and -8A:T were modeled as -9A:T and -8G:C. A multistate design calculation [15] was performed to select amino acids at positions 24Z, 26Z, 27Z, and 29Z. Three states were included in the design. The first state was the target state, which was modeled using the altered DNA structure. The second state was the original structure with the wild-type DNA sequence and served as a competitor to enforce binding specificity of the selected proteins for the altered recognition site (negative design state). The third state was the modeled structure of the best single-state design for the target state with the modified DNA sequence, and the energy associated with this state is a constant during the multi-state design procedure. It represents the best scoring

protein-altered DNA complex as assessed with the Rosetta energy potential, and it is therefore impossible for the energy associated with the target state to be lower than this value. As a result, multiple calculations were performed which differed from each other only in an artificial offset applied to the third state. Progressively larger offsets bias the calculations to select sequences that achieve higher specificity for the first state over the second state at the expense of achieving Rosetta scores that are allowed to be progressively worse than the third state.

A genetic algorithm was used to evolve a population of sequences that prefer the target state to the two competitors. An initial population of 2000 sequences was generated by selecting random amino acids at the four design positions. The side chain conformations of these four residues (with the rest of the protein and DNA structure held fixed) were predicted for the first and second states using a Monte Carlo algorithm, and the Rosetta score recorded. As noted above, the energy of the third state is a constant. A 'fitness' score for each sequence *i* in the population is calculated:

$$\text{Fitness}_i = E_{target} - \langle E_{competitors} \rangle$$

Where E_{target} is the energy of the target state, and the brackets denote an ensemble (Boltzmann weighted) average over the energies of the competitors. Conceptually, the fitness corresponds to the transfer free energy of the protein from the ensemble of competitors to the target state. Subsequent generations were constructed using the following procedure. First, the sequence with the best (lowest) fitness was promoted automatically. Next 1980 sequences were created by recombining two members of the population using uniform crossover of two parents chosen by tournament selection [16]. Finally, the remaining 19 sequences were generated by mutating a single parent chosen by tournament selection with a 25% chance of randomizing each position in turn. A fitness value was calculated for each new sequence, and the population was propagated for 30 generations.

Appendix II.2 Methods for Chapters 3 and 4

Computational tools

All protocols were implemented within the Rosetta molecular modeling package, and will be available for free academic use through the Rosetta Commons. They are currently available to institutions participating in Rosetta Commons (or upon request), and the code revision numbers are 44353 for trunk Rosetta and 44354 for the version with the energy function optimized here and the DNA-Rebuild method (source/workspaces/blab/mini). The energy function was similarly optimized for trunk Rosetta, however the orientation-dependent desolvation is not available and the reference energies differed (Figure AI.12). These two code versions and energy functions will be integrated in a future release of Rosetta. The executables currently available in both code versions are *dna_motif_collector* for the generation of motif libraries and *motif_dna_packer_design* for designing with a motif bias. The flexible DNA simulations are currently limited to the workspaces branch and the executable that rebuilds the DNA and designs with motifs is called *dna_fragment_rebuild_with_motifs*. The designs completed with an increased temperature for the low temperature of the simulated annealing algorithm and removal of the final quenching step for the packer are based in the *motif_dna_packer_design*, but require the modification of two lines prior to compilation. These changes are detailed in the supplemental. An additional executable, *failure_analyzer*, for analysis of the design data (failure identification, energy differences between designs) is available in a later revision (source/workspaces/blab/mini, revision 45873). Many parameters of all methods are modifiable via the command line and all currently available options are discussed in the supplemental methods. Other data available upon request includes, but is not limited to, the final list of PDB codes used to generate the library, the complete motif library in either a single file or in the form of two-residue PDB files, and python analysis scripts (also available in /source/workspaces/sthyme/scripts).

Structural data for training and test set

A set of 112 largely non-redundant, crystallized protein-DNA complexes all with a resolution of lower than 2.5 Å was downloaded from the RCSB protein data bank [63]. This set split into one group of 48 complexes and another of 64 complexes; the group containing 48 pdbs was used for training the energy function and the group containing 64 was used for testing and analyzing improvements identified from the training procedure. All pdbs were downloaded as the biological assemblies, and several required small modifications for compatibility with the subsequent Rosetta protocols and analysis scripts.

Training set: 1a1f, 1a3q, 1az0, 1bc8, 1bdt, 1bl0, 1ckq, 1d02, 1dc1, 1e3o, 1f4k, 1gd2, 1gu4, 1hcq, 1iaw, 1ig7, 1ign, 1j1v, 1jnm, 1lmb, 1lq1, 1m5x, 1mjo, 1mnm, 1mnn, 1nkp, 1ozj, 1pp7, 1puf, 1r4o, 1r71, 1r7m, 1skn, 1tc3, 1ubc, 1w0u, 1wte, 1zs4, 2bam, 2d5v, 2ex5, 2ezv, 2fl3, 2h27, 2hdd, 2oaa, 2qoj, 3pvi

Test set: 1a1h, 1a73, 1aay, 1am9, 1b3t, 1b94, 1dfm, 1dmu, 1dp7, 1egw, 1g2f, 1g9y, 1hcr, 1hwt, 1i3j, 1jey, 1jft, 1k61, 1mey, 1mow, 1mus, 1nvp, 1oe5, 1oup, 1qpi, 1r0o, 1sa3, 1tup, 1xbr, 2bop 2c9l, 2dgc, 2e52, 2fqz, 2o4a, 2odi, 2or1, 2wt7, 2x6v, 2xqc, 2xsd, 2z3x, 3bm3, 3bs1, 3c25, 2co6, 3fc3, 2fdq, 3h0d, 3iag, 3igm, 3jtg, 3jxb, 3jy1, 3lnq, 3m4a, 3mln, 3mqy, 3mx4, 3n7q, 3o9x, 3pvv, 3qqy, 6pax

Generation of motif library

A motif is defined as the spatial arrangement of six atoms. In the case of a protein-DNA motif, three of these atoms are located on a DNA base that interacts with a protein residue and the other three are derived from that protein residue (Figure 11). This geometric relationship is expressed as a translation vector and a set of Euler angles, as previously described [17]. The atoms that define motifs are currently fixed for different amino acid and DNA residues. Motifs were collected from protein-DNA complexes with a resolution of better than 2.8 Å that were downloaded from the RCSB protein data bank on August 9th, 2011. The set initially consisted of 1459 complexes, which was reduced to 1375 complexes

after removal of PDBs that were not compatible with Rosetta without manipulation of the PDB files or modification of Rosetta.

The motif library used for this work includes both major and minor groove interactions, as well as water-mediated contacts. The collection algorithm is defined by iteration over every protein residue in each of the protein-DNA complexes and the identification of up to two DNA bases that have the greatest amount of Rosetta interaction energy with that protein residue. This interaction energy between the protein and DNA residue is defined as a packing score (combined attractive and repulsive energies), a direct sidechain-sidechain hydrogen bonding score, and a water-mediated hydrogen bonding score, if a theoretical water can be placed at a canonical location on the DNA base [18]. To count as a motif interaction, the protein-DNA pair must have either a packing score of less than -0.5 Rosetta energy units (REUs), a direct hydrogen bonding score of less than -0.3 REUs, or a water-mediated hydrogen bonding score of less than -0.3 REUs.

Redundancy in the motif library arises mainly from the inclusion of multiple crystal structures of the protein-DNA complex or from equivalent monomers of homo-oligomeric complexes. To reduce this redundancy, the amino acid and DNA residue pairs are all placed in the same coordinate frame, based around the motif atoms of the DNA base, for all interactions involving that type of DNA residue. Any DNA residue that has less than 0.2 RMSD over the heavy atoms with any other DNA residue is eliminated from the motif library.

Removal of homologous motifs from the motif library

Prior to identifying motif interactions that can be made in a particular protein-DNA complex, it is necessary to remove motifs derived from that same PDB entry or from one of a homologous protein. The inclusion of such motifs would result in artificial biases toward the native sequence. The protocol developed for the remove consists of a BLAST [19] run against the PDB database identified all structures with an e-value of less than 0.05 to the starting structure and a python script to parse the output of the blast run and remove homologous motifs from the library.

Identification of rotamers forming motif interactions

The utilization of motifs in fixed-backbone protein design requires the identification of amino acid rotamers that are capable of forming a motif interaction in a given protein-DNA complex. To accomplish this goal, backbone-dependent rotamers derived from the Dunbrack rotamer library [20], included with the ROSETTA software, are built at protein positions in a protein-DNA interface. Interface positions are identified using a previously described protocol [11] that builds a set of arginine rotamers at each protein position and checks whether any nucleotide base atom is within 3.8 Å of these arginine sidechains. For this motif search protocol, the level of rotamer sampling was set to include extra sampling at χ1-4, as well as an additional four half-step deviations from the bin of the rotamer. Each rotamer is screened against all nearby DNA bases to test whether a motif interaction can be made and it must pass several cutoffs to be considered a successful rotamer. First, a single atom from a canonical DNA base defined by the motif being tested, currently the C1*, is placed via the defined motif orientation. A distance between this atom and every nearby C1* in the crystal structure DNA is calculated. Passing a defined distance cutoff, set to be 2.0 Å for these experiments, allows the rotamer to be subject to further testing. The next test screens for how parallel a motif-placed canonical base is to the closest crystal structure base by the calculation of a dot product for vectors perpendicular to the plane of the 6 atoms of a placed nucleobase and the crystal structure nucleobase. The dot product for these experiments was set to be greater than 0.97 to be considered for a final test of the rmsd over the same 6 atoms of the nucleobase compared with the nearby crystal structure nucleobase. For these experiments, the rmsd had to be less than 1.0 in order for the rotamer to be able to make a successful motif contact. Both the distance and rmsd cutoffs are automatically reduced for motifs with longer sidechains that have many more rotamers. Cutoffs for arginine are cut two-fold and cutoffs for methionine, lysine, glutamate, and glutamine are cut by a third. All rotamers passing the cutoffs are then sorted, dependent on a combined score of the rmsd and dot product (rmsd divided by dot product), and the lowest scored rotamers are preferentially considered to be successful if the user indicates a limit on the number of rotamers to be utilized by further design protocols. The default limit is set to be 100 rotamers of each amino

acid type at each protein position being designed and this default was maintained in the experiments described here.

Motif-biased design

Rotamers identified to make motif interactions with the search procedure described in the preceding section are incorporated into the standard design procedure by adding them to the rotamer set being used by the packer. For these experiments, the initial rotamer set included extra sampling of $\chi 1$ and $\chi 2$, and three 1/3 step additional deviation samples for $\chi 1$ and $\chi 2$ of aromatic residues. The packer provides the core functionality for Rosetta design, utilizing a Monte Carlo simulated annealing algorithm, guided by a physically based atomic-level forcefield[60]. These motif rotamers are flagged and can be given an energy bonus over other rotamers in the rotamer set. The flag is implemented as a residue patch called SpecialRotamer and the energy term special_rot allows for the user to implement differential bonuses for these rotamers. Alternatively, there are input options that support the definition of a starting motif bonus and a subsequent number of steps of two-fold reduction of that bonus, producing multiple designs each with a different bias toward inclusion of these rotamers. The designs completed in this work cover the range of bonuses from -10 to -1.25. Additional designs where motif rotamers are added with no weight and where motif rotamers are left out of the rotamer set are produced by default. Identification of protein positions where mutation of the protein sequence is allowed is described in the section on collection of motif rotamers, as it occurs by the same method. An additional shell of residues surrounding these designable residues are allowed to change rotamer conformation, but not protein sequence.

For the sequence recovery work individual design runs were done at every single base-pair in the interface, simulating the approach used for specificity redesign where only a small group of amino acids are designed simultaneously. Energy function analysis and optimization was guided by sequence recovery calculations. Two metrics, weighted and unweighted recovery, were calculated for each set of design calculations. The unweighted metric counts every designed position equally, and the weighted metric is an average over the recoveries for each amino acid type and free from biases in the amino acid composition of

the interface positions. The inclusion of the weighted metric during optimization is necessary to avoid artificial improvements in overall recovery due to biasing the energy function toward recovery of amino acids that are overrepresented in protein-DNA interfaces, namely lysine and arginine, at the expense of the less abundant types. A previously improved weight set [21] that was optimized without consideration of the weighted metric contains this particular bias (Figure AI.6).

Flexible DNA interface design

Use of the flexible DNA interface design protocol was limited to computationally tractable PDBs that were compatible with the DNA movement portions of the protocol without any modification or reformatting. This method consists of a previously described [22] DNA rebuilding step followed by a Motif-biased design run. For each targeted DNA design, that base-pair and the two surrounding base-pairs were allowed to move. Unpaired DNA base-pairs, DNA strands containing chain internal chain breaks, or base-pairs on the end or one away from the end of DNA chain were not included because they are not compatible with the DNA rebuilding portion of the protocol. After each design calculation, the rebuilt DNA was allowed to minimize prior to the next design iteration (between each round of lowering the motif bonus).

Rebuild set: 1a1f, 1a1h, 1a3q, 1aay, 1az0, 1bc8, 1bdt, 1bl0, 1ckq, 1d02, 1dc1, 1e3o, 1egw, 1f4k, 1g2f, 1gd2, 1gu4, 1hcq, 1hwt, 1i3j, 1ig7, 1ign, 1j1v, 1jnm, 1lq1, 1m5x, 1mey, 1mnm, 1mnn, 1nkp, 1oe5, 1ozj, 1pp7, 1puf, 1r0o, 1r71, 1r7m, 1sa3, 1skn, 1tc3, 1ubd, 1w0u, 1wte, 1xbr, 1zs4, 2bam, 2c9l, 2d5v, 2e52, 2ex5, 2ezv, 2fl3, 2h27, 2hdd, 2o4a, 2oaa, 2qoj, 2wt7, 2xsd, 2z3x, 3c25, 2co6, 3fc3, 3fdq, 3h0d, 3iag, 3jtg, 3jxb, 3lnq, 3m4a, 3mln, 3mx4, 3n7q, 3o9x, 3pvi, 3pvv, 3qqy, 6pax

Identification of failed design pockets

The metrics designating an incorrectly designed position as not being a true failure are as follows: 1) The correct amino acid type being seen for over 25% of the design runs from the

set of designs completed with a varying motif weight, indicating that the wild-type is favorable in the context of a motif bonus. 2) The wild-type amino acid making very little contact to any protein or DNA residue, as defined by a total Rosetta interaction energy with all nearby residues of no more than -2 REUs. 3) The wild-type amino acid being one of the smallest amino acids types, because native protein-DNA interfaces are not always optimized for the tight binding and high specificity that the computational methods are programmed to produce and a small amino acid type being redesigned to a larger one with more contacts is potentially an acceptable change that could increase interface affinity. 4) The designed amino acid being chemically related to the wild-type amino acid and likely to be making a similar contact, such as a glutamate being redesigned to a glutamine. Future implementations could utilize atom-type specific analyses for a more accurate assessment of contact success.

Bacterial screen

A bacterial screen for active variants of I-AniI was completed as previously described [23], albeit with minor modifications. Electrocompetent *E. Coli* cells, the DH12S strain from Invitrogen, were transformed with a pCCDb plasmid containing two adjacent copies of the I-AniI LIB4 target site [24], a variant of the wild-type target site containing 2 activating substitutions. This pCCDb containing strain was prepared for the selection using a standard procedure for electrocompetent cell preparation. Each of the 44 libraries, corresponding to the 44 interface positions, were ligated and the pCCDb containing electrocompetent cells were transformed with the purified ligation products. Transformants were recovered in TB media for a half hour at 37ºC. The selection procedure was completed for 4 hours in 2 mL liquid culture at 30ºC. Following liquid selection, 1 µL was plated on each of minimal selection (100 µg/mL carbenicillin, 1 mM IPTG, and 0.02% L-arabinose) and control (100 µg/mL carbenicillin) plates (1.5% agar, M9 salt, 1% glycerol, 0.8% tryptone, 0.2% thiamine, 1 mM $MgSO_4$, 1 mM $CaCl_2$) and grown for ca. 36 hours at 30ºC. Approximately 20 colonies were picked from each selection plate for each of the 44 positions, grown overnight in 96-well culture plates, and submitted for sequencing as 96-well plate glycerol stocks to the GENEWIZ sequencing facility.

Construction of plasmids and libraries

The pCCDb plasmid containing the I-AniI LIB4 [24] target sites was built by phosphorylating and annealing oligonucleotides from Integrated DNA Technologies to form a duplex with sticky ends compatible with the NheI and SacII restriction sites in the pCCDb vector [23]. An amino acid library was built for each of the 44 protein interface positions, using assembly PCR [2] with oligonucleotides containing an NNS codon (Integrated DNA Technologies) at the randomized position. These libraries were ligated into pEndo vector [23] between the NcoI and NotI restriction sites and screened for activity in the bacterial selection system. All C-terminal I-AniI libraries (starting at position 148) were built in the context of the activating M5 [25] mutations, and all N-terminal mutations (from position 18 to position 72) were built in the context of M4, which is M5 without the I55V mutation.

Detailed computational methods

Specific modifications to the ROSETTA energy function

Phosphorous Desolvation
The LK_DGFREE term in the atom_properties.txt file of the database is changed from -24 to -4.1 and the LK_VOLUME term is changed from to 34.8 to 14.7.
Electrostatics
The standard.wts file for the Standard ROSETTA energy function was modified by the replacement of fa_pair term (weight of 0.49) with the hack_elec term (weight of 0.5).
LK_Ball
The Electrostatics energy function was modified by the addition of several energy terms corresponding to the orientation-dependent desolvation method (these terms replace fa_sol 0.65):

lk_ball 0.325, lk_polar 0.325, lk_polar_nw 0.65, lk_nonpolar 0.65, lk_charged 0.325, lk_ball_xd 0.325, lk_polar_xd 0.325, lk_polar_nw_xd 0.65, lk_nonpolar_xd 0.65, lk_ball_dd 0.5, lk_polar_nw_dd 0.5, lk_nonpolar_dd 0.65

Multi-Desolvation
The LK_DGFREE term was modified for the following atom types in the atom_properties.txt file:

NH20 from -10 to -7.8 (modifying Gln and Asn)
ONH2 from -10 to -5.85 (modifying Gln and Asn)
Nlys from -20 to -16 (modifying Lys)
Narg from -11 to -10 (modifying Arg)

Corresponding reference energy changes:

Asp from -0.67 to -0.75

Glu from -0.81 to -0.71
Lys from -0.65 to -1.2
Asn from -0.89 to -0.8
Gln from -0.97 to -0.78

Attractive

The fa_atr term was increased from 0.8 to 0.95. Corresponding reference energy changes:

Phe from 0.63 to 1.63
Trp from 0.91 to 2.21
Tyr from 0.51 to 0.91

Lysine Charge

The positive charge on the hydrogens on the terminal nitrogen of lysine was increased from 0.33 to 0.48.

Reference Energies

The following reference energies were modified:

Ala from 0.16 to 0.26
Cys from 1.7 to 0.5
Phe from 1.63 to 1.7
His from 0.56 to 0.8
Ile from 0.24 to -0.1
Met from -0.34 to -0.1
Arg from -0.98 to -0.65
Ser from -0.37 to -0.57
Thr from -0.27 to -0.8
Val from 0.29 to -0.1
Trp from 2.21 to 2.3
Tyr from 0.91 to 1.0

Additional changes, essentially negligible for sequence recovery, for this "Optimized" energy function include the change of hbond_sc from 1.1 to 1.17 and the addition of the two command line options "-*local_bb_sc_downweight* 0.2" and "-*apply_proton_chi_potential*".

Code modification for the HighTemp-Packer calculation

The lower temperature of the simulated annealing algorithm was changed from 0.3 to 1.3 because it provided a similar level of diversity to the diversity derived from the DNA-Rebuild method. The final quenching step of the simulated annealing algorithm is also removed. The two lines of code for these changes are in the /mini/src/core/annealer/SimAnnealerBase.cc.

Original Line 44: const PackerEnergy SimAnnealerBase::lowtemp = 0.3;
New Line 44: const PackerEnergy SimAnnealerBase::lowtemp = 1.3;
Original Line 272: void SimAnnealerBase::set_to_quench(){ quench_ = true;}
New Line 272: void SimAnnealerBase::set_to_quench(){ quench_ = false;}

Available command-line options for motif Protocols

Used only with dna_motif_collector for collection of the motif library.

'keep_motif_xtal_location', 'Boolean', default = 'false', desc= 'controls whether motifs are

moved away from original PDB location (comparison between motifs is easier if they are moved, so that's default).'
'pack_score_cutoff', 'Real', default = '-0.5', desc = 'fa_atr (attractive) + fa_rep (repulsive) energy threshold for a two-residue interaction to determine if it is a motif.'
'hb_score_cutoff', 'Real', default = '-0.3', desc = 'hbond_sc (sidechain-sidechain hydrogen bonding) energy threshold for a two-residue interaction to determine if it is a motif.'
'water_score_cutoff', 'Real', default = '-0.3', desc = 'h2o_hbond (water hydrogen bonding) energy threshold for a two-residue interaction to determine if it is a motif.'
'motif_output_directory', 'String', desc = 'path for the directory where all the motifs are collected as 2-residue pdbs.'
'eliminate_weak_motifs', 'Boolean', default = 'true', desc= 'controls whether only the top 1-2 motifs (instead of all possible interactions) are counted for every protein position in a protein-DNA interface.'
'duplicate_motif_cutoff', 'Real', default = '0.2', desc = 'RMSD cutoff for an identical, canonical base residue placed via a motif to see if that motif already exists in a motif library.'
'preminimize_motif_pdbs', 'Boolean', default = 'false', desc= 'controls whether the input PDB structure sidechains and backbone are minimized before motifs are collected.'
'preminimize_motif_pdbs_sconly', 'Boolean', default = 'false', desc= 'controls whether the input PDB structure sidechains are minimized before motifs are collected.'
'place_adduct_waters', 'Boolean', default = 'true', desc= 'whether or not adduct waters are placed before motifs are collected, there will be no water interaction energy calculated if this option is false.'

Example for dna_motif_collector:

/rosetta/bin/dna_motifs_collector.linuxgccrelease *-motif_output_directory* Motif_Dir_August2011/ *-ignore_unrecognized_res -adducts* dna_major_groove_water *-database* /minirosetta_database/ *-l* list_PDBsAugust2011

> ***'ignore_unrecognized_res'** and *'database'* are standard ROSETTA options that should be included in all commandlines and are not specific to motif protocols
> ***'adducts'** is an option that can be used with non-motif protocols, but it is specific to DNA. The specification of 'dna_major_groove_water' results in the potential placement of adduct waters at canonical DNA major groove positions. The motif option *'place_adduct_waters'* will not function without this option.

Used with both motif_dna_packer_design and dna_fragment_rebuild_with_motifs.
'list_motifs', 'FileVector', desc= 'File(s) containing list(s) of two-residue, motif PDB files to process and use as the motif library.'
'motif_filename', 'String', desc= 'File containing motifs – can be used to build the motif library as an alternative to the -list_motifs that takes PDB files.'
'BPData', 'String', desc= 'File containing BuildPosition (designed protein position) specific motifs and/or rotamers.'
'list_dnaconformers', 'FileVector', desc= 'File(s) containing list(s) of PDB files to process. This option is included to allow the use of DNA residues collected from the PDB instead of a canonical DNA residue to determine if a motif passes required *z2* and *r2* cutoffs during the

motif search. This option was not used during the work described in this paper.'
'target_dna_defs', 'StringVector', default = '', desc = 'Can be used in conjunction with the standard way of making DNA mutations/designable areas (*dna::design::dna_defs*) to specifically target the motif searching to a subset of the mutated bases or to do motif searches with multiple allowed DNA base types (for DNA target site prediction). The form is the same as for the standard *dna_def* – chain.position.type or X.409.ADE.'
'motif_build_defs', 'StringVector', default = '', desc = 'The protein positions that can be searched for motifs. This option is most useful to limit the amino acid types that can be included in the motif search. The format is same as for the dna_def, except that one letter codes are used – Z.33.SR would allow motifs of serine or arginine at position Z33.'
'r1', 'Real', default = '4.5', lower = '0', desc = 'RMSD cutoff between motif anchor position and motif target position for allowing a particular motif rotamer to continue on to expand with DNA conformers or just to pass with the canonical base. The RMSD is calculated between a canonical base and the nearest crystal structure base of the same type. Six atoms in the plane of the nucleobase portion of the DNA residues are used for the RMSD calculation.'
'r2', 'Real', default = '1.1', lower = '0', desc = 'RMSD cutoff between motif anchor position and motif target position for accepting the motif. The RMSD is calculated between a canonical base and the nearest crystal structure base of the same type. Six atoms in the plane of the nucleobase portion of the DNA residues are used for the RMSD calculation.'
'z1', 'Real', default = '0.75', lower = '0', desc = 'DNA motif specific: cutoff between motif target DNA position and canonical base for allowing a particular motif to continue on to expand with DNA conformers. This cutoff is a test for how parallel the canonical, placed base is to the nearest crystal structure base (dot product of two vectors from the base plane).'
'z2', 'Real', default = '0.95', lower = '0', desc = 'DNA motif specific: cutoff between motif target DNA position and DNA conformer, or repeated with a canonical base and equivalent to *z1*, placed according to motif for accepting the pair of residues. This cutoff is a test for how parallel the canonical, placed base is to the nearest crystal structure base (dot product of two vectors from the base plane).'
'dtest', 'Real', default = '5.5', lower = '0', desc = 'DNA motif specific: cutoff between motif target DNA position and DNA conformer or canonical DNA base placed according to motif for accepting the pair of residues. This cutoff is based on the C1* atom that is found in all DNA residues.'
'rotlevel', 'Integer', default = '5', lower = '1', desc = 'level of rotamer sampling for motif search. The recommended sampling is 6, but 8 (slower and longer motif search runs) was used for the data collected in this paper.'
'num_repacks', 'Integer', default = '5', lower = '0', desc = 'number of cycles of dropping special_rot weight and design.'
'minimize', 'Boolean', default = 'true', desc = 'whether or not to minimize the motif rotamers that pass the cutoffs toward the xtal structure DNA (via the constraint of the motif interaction) prior to adding them to the packer.'
'run_motifs', 'Boolean', default = 'true', desc = 'whether or not to collect and include motif rotamers in design.'
'expand_motifs', 'Boolean', default = 'true', desc = 'whether or not to expand motifs by generating sets of designs for each type of motif (by amino acid type) identified in the search step. This option is used to generate the motif-based sequence constraint data.'
'aromatic_motifs', 'Boolean', default = 'false', desc = 'whether or not to use expand motifs,

using aromatic motifs only.'
'dump_motifs', 'Boolean', default = 'false', desc = 'whether or not to output pdbs with the best rotamer/conformer for each motifs.'
'quick_and_dirty', 'Boolean', default = 'false', desc = 'quick motif run to get a list of all possible motifs before doing a real run. This type of run stops after the *dtest*, *z1*, and *r1* tests.'
'special_rotweight', 'Real', default = '-40.0', desc = 'starting weight for the weight on motif rotamers.'
'output_file', 'String', desc = 'name of output file for all the best motifs and rotamers or for the dna_motif_collector it is the file where all the motifs are dumped.'
'data_file', 'String', desc = 'name of output file for any data about how many rotamers and motifs pass what tests, etc.'
'clear_bprots', 'Boolean', default = 'false', desc = 'whether or not to clear the rotamers that were read in from a previous run and restart with only the motifs that were read in and the specified rotlevel.'
'rots2add', 'Integer', default = '100', lower = '1', desc = 'number of rotamers to add to the packer from the motif search for each amino acid type. Too many or all passing rotamers will is not allowable due to computational memory constraints.'
'restrict_to_wt', 'Boolean', default = 'true', desc = 'restrict the motif search to finding motifs of the same amino acid as the starting pose, such as for homology modeling applications.'
'rerun_motifsearch', 'Boolean', default = 'true', desc = 'setting the motif search process to run again, using the rotamers in the build position, most likely to change stringency or allowed amino acid type on a second run.'

Example for motif_dna_packer_design:

/rosetta/bin/motif_dna_packer_design.linuxgccrelease *-run_motifs -dtest* 2.0 *-z1* 0.97 *-z2* 0.97 *-r1* 1.0 *-r2* 1.0 *-dna::design::z_cutoff* 6.0 *-motifs::rotlevel* 8 *-motifs::list_motifs* ../list_August2011Motifs_2QOJ *-motifs::output_file* ./2QOJ.motifs *-s* ../2QOJ.pdb *-score::weights* /work/sthyme/weights/dna_march2011_sr.wts *-ignore_unrecognized_res -database* /work/sthyme/sparse_databases/minirosetta_database_sparse_JA_march2011/ /minirosetta_database/ *-ex1 -ex2 -ex1aro::level* 6 *-ex2aro::level* 6 *-extrachi_cutoff* 0 *-dna::design::dna_defs* X.409.CYT *-special_rotweight* -20.0 *-num_repacks* 4

*The motif search specific options shown here match the options used for all design calculations described in this work.
**'ignore_unrecognized_res'* and *'database'* are standard ROSETTA options that should be included in all commandlines and are not specific to motif protocols
**'score::weights'* is a standard ROSETTA option that allows the energy function to be externally chosen.
**'ex1', 'ex2', 'ex1aro::level', 'ex2aro::level',* and *'extrachi_cutoff'* are standard ROSETTA options that define the rotamer set used by the standard packer. The *'level'* specification for *'ex1aro'* and *'ex2aro'* refers to extra standard deviation sampling for $\chi 1$ and $\chi 2$ of aromatic residues. This level of rotamer sampling was used for all design calculations described in this work. The extra sampling for the aromatic residues was chosen because these residues are likely to be mis-designed if there is no

available rotamer that is close to the native aromatic due to repulsive forces.
**'dna::design::z_cutoff'* and *'dna::design::dna_defs'* are specific to DNA design methods, where the *'z_cutoff'* determines the designable residues and the *'dna_defs'* provide a way to mutate DNA positions and/or identify a designable area. The larger the input value to the *'z_cutoff'*, the more residues surrounding the input bases to *'dna_defs'* will be considered designable (6.0 was used for all the work described here).

Currently only available in /workspace/blab version of ROSETTA.
'minimize_dna', 'Boolean', default = 'true', desc = 'whether or not to minimize DNA after every round of design with special_rot weight dropping.'
'flex_sugar', 'Boolean', default = 'false', desc = 'whether or not to add flexibility to the DNA sugar pucker.'
'dna_geometry', 'Boolean', default = 'false', desc = 'show DNA geometry calculations, only relevant for the dna_fragment_rebuild_with_motifs protocol.'
'apply_proton_chi_potential' #Used with "Optimized" energy function, but has negligible effect on sequence recovery.
*'local_bb_sc_downweight' '*0.2*'* #Used with "Optimized" energy function, but has negligible effect on sequence recovery.

Example for dna_fragment_rebuild_with_motifs:
Same as for the motif_dna_packer_design with the following additions:
-minimize_dna
-double_frag_vall ../double_fragment_vall_v2.lib
**'dna::design::dna_defs'* should always be included for rebuilding with input DNA positions. The rebuilding will occur with the input dna_def position and the two surrounding bases, provided that none of the bases to be rebuilt are at termini or breaks in the DNA chain.

Appendix II.3 Methods for Chapter 5

Identification of homologues and target sites

I-AniI homologues were identified by searches against the NCBI non-redundant (nr) database with BLAST – blastp and tblastn – using the I-AniI ORF as the query sequence [26, 27]. Target sites were identified by examination of the nucleotide sequence on either side of the intron containing the homologue ORF. The high similarity of the putative target sites to the I-AniI target supported these predictions. Multiple sequence alignments of both the homologues and their predicted target sites were constructed using Jalview (Figure 20) [28].

These putative endonucleases are denoted by the suffix P in accordance with nomenclature guidelines [29].

Generation of hybrid I-AniI homologue endonucleases

The amino acid sequence of each homologue of I-AniI was modeled onto the crystal structure of the wild-type I-AniI protein (2QOJ, [24]). The I-AniI protein scaffold used for generating hybrid endonucleases contained point mutations that have been shown to increase either solubility or cleavage activity at physiological temperatures [24.25]. All C-terminal pocket transfers were made in the context of the activating mutation F13Y, while all N-terminal transfers, with the exception of the K24N/T29K enzyme pair and two variants derived from the I-VinIP homologue, were made in the context of S111Y ("Base Activity" column, Table 1). These two activating mutations were identified [25] during the course of this study and the choice of which activating mutations to include for the each hybrid protein was a function of experimental timing, the maintenance of consistency for C- and N-terminal transfers, and considerations regarding positions of the activating mutation in relationship to the transferred pocket. In all cases, the hybrids were compared to I-AniI with the corresponding activating mutation for consistency. See Table S2 for the complete list of mutations made for each hybrid.

Expression and purification of proteins

Genes for each homologue-based variant of I-AniI were assembled, cloned, sequence-verified, and transformed into BL21 Star cells (Invitrogen). A half-liter or one liter culture of auto-induction media [3] was then inoculated and grown at 37°C for 8 -12 hours until approximate saturation, after which expression at 18°C was allowed for 20 - 24 hours. Cells were then harvested and resuspended in 20mM Tris pH 7.5, 30 mM Imidazole, and 1.0M NaCl prior to lysis via a freeze-thaw cycle, sonication, and the addition of lysozyme.

Proteins were isolated from the soluble fraction with nickel affinity chromatography. The purified proteins were concentrated, buffer exchanged in 20mM Tris pH 7.5 and 500mM NaCl, and stored in 50% (v/v) glycerol. Purity of ca. >95% for all samples was confirmed by

SDS-PAGE and the mass and purity were additionally verified by mass spectrometry. Protein concentration for each sample was determined by measuring absorbance at 280 nm.

DNA cleavage assays

Plasmid DNA substrates, containing single base-pair substitutions from the I-AniI wild-type target site, were constructed by site-directed mutagenesis according to methods described elsewhere [6] and linearized with the restriction endonuclease ScaI. For optimized enzyme activity and stability, the reaction buffer contained final concentrations of 170 mM KCl, 10 mM MgCl2, and 20 mM Tris pH 9.0. For every I-AniI variant assayed, 8 serial 2-fold dilutions of enzyme were performed in 1.25X reaction buffer, ranging from 5 nM to 1500 nM depending on the experiment, and each dilution was incubated with 100 ng (ca. 5 nM) of linearized substrate for 30 minutes at 37°C. Wild-type I-AniI, containing the same activating mutations as each hybrid, was tested in parallel under the same reaction conditions. Reactions were quenched with ca. 17 nM EDTA, followed by 60°C incubation for 5-10 minutes. The resulting cleavage products were separated by gel electrophoresis on 1.2% agarose TBE gel and were visualized by staining with ethidium bromide.

The data were analyzed as previously described by quantifying the spectral density of substrate and product bands using ImageJ. The percent cleavage was calculated by dividing the sum of the two product band densities by the sum of the densities of all three bands and was plotted versus enzyme concentration in GraphPad Prism. At least two independent determinations of each enzyme cleavage profile were performed and inspected for substrate degradation and experimental error prior to being reported here.

In order to estimate the concentrations (nM) of enzyme corresponding to half-maximal cleavage of the target site ($EC_{0.5max}$), data were fit to a sigmoid function as follows:

$$f_{([endonuclease])} = \frac{f_{\max} * [endonuclease]^h}{EC_{0.5\max} + [endonuclease]^h}$$

Here, $f_{([endonuclease])}$ is the fraction of DNA site cleavage corresponding to endonuclease concentration in nM as denoted by [*endonuclease*]. f_{max} is the maximal fraction of site cleavage, with 1 being its greatest allowable value indicating complete cleavage of the

substrate with no remaining uncut fraction. While the value of the Hill coefficient *h* is 1 for a simple hyperbolic formula, setting *h* to 1.5 allowed a sigmoid function that consistently achieved a better fit to the data. $EC_{0.5max}$ could then be determined by solving the equation.

II.4 Appendix II Bibliography

[1] Bolduc, J. M. *et al.* Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.* **17**, 2875-2888 (2003)
[2] Stemmer, W. P. C., Crameri, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49-53 (1995)
[3] Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005)
[4] Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995)
[5] http://rsbweb.nih.gov/ij/
[6] Kunkel, T. A., Roberts, J. D. & Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol.* **154**, 367–382 (1987)
[7] Geese, W. J., Kwon, Y. K. & Waring, R. B. *In vitro* analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-AniI. Eur. J. Biochem. **270**, 1543-1554 (2003)
[8] Halford, S. E., Johnson, N. P., & Grinsted, J. The EcoRI restriction endonuclease with bacteriophage lambda DNA. Kinetic studies. *Biochem. J.* **191**, 581-592 (1980)
[9] Zhao, L., Pellenz, S. & Stoddard, B. L. Activity and specificity of the bacterial PD-(D/E)XK homing endonuclease I-*Ssp6803*I. *J. Mol. Biol.* **385**, 1498-1510 (2008)
[10] Havranek, J. J., Duarte, C. M. & Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.* **344**, 59-70 (2004)
[11] Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656-659 (2006)
[12] Canutescu, A. A. & Dunbrack, R. L., Jr. Cyclic coordinate descent: robotic algorithm for protein loop closure. *Protein Sci.* **12**, 963-972 (2003)
[13] Das, R. *et. al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**, 118-128 (2007)
[14] Ashworth J. & Baker, D. Assessment of optimization of affinity and specificity at protein-DNA interfaces. *Nucleic Acids Res.* **37**, e73 (2009)
[15] Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10,** 45-52 (2003)
[16] Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.
[17] Havranek, J. J. & Baker, D. Motif-directed flexible backbone design of functional interactions. *Protein Sci.* **18**, 1293-205 (2009)
[18] Jiang, L., Kuhlman, B., Kortemme, T., & Baker, D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **58**, 893-904 (2005)
[19] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
[20] Dunbrack, R. L. Jr. & Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661-8 (1997)

[21] Ashworth, J., Taylor, G. K., Havranek, J. J., Quadri, S. A., Stoddard, B. L., & Baker, D. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.* **38**, 5601-5608 (2010)
[22] Yanover, C. & Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* **39**, 4564-4576 (2011)
[23] Doyon, J. B., Pattanayak, V., Meyer, C. B. & Liu, D. R. Directed evolution and substrate specificity profiling of homing endonuclease I-SceI. *J. Am. Chem. Soc.* **128**, 2477-84 (2006)
[24] Scalley-Kim, M., McConnell-Smith, A. & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**, 1305-1319 (2007)
[25] Takeuchi, R., Certo, M., Caprara, M. G., Scharenberg, A. M., & Stoddard, B. L. Optimization of *in vivo* activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation. *Nucleic Acids Res.* **37**, 877-890 (2008)
[26] http://www.ncbi.nlm.nih.gov/blast/blast_databases.shtml
[27] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.* **215**, 403-410
[28] Clamp, M., Cuff, J., Searle, S.M., Barton, G.J (2004) The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427
[29] Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S.Kh., Dryden, D.T., Dybvig, K., et al. (2003) *Nucleic Acids Res.* **31**, 1805-1812

Vita

Summer Thyme was born in Keene, New Hampshire. She attended Scripps College in Claremont, California from 2002 to 2006. At Scripps she earned a Bachelor of Arts degree in Biology and a second one in Chemistry. In 2012 she earned a Doctor of Philosophy at the University of Washington in Biochemistry.