Engineering proteins that bind, move, make and break DNA
Cynthia H Collins*, Yohei Yokobayashi†, Daisuke Umeno† and Frances H Arnold†‡

Recent protein engineering efforts have generated artificial transcription factors that bind new target DNA sequences and enzymes that modify DNA at new target sites. Zinc-finger-based transcription factors are favored targets for design; important technological advances in their construction and numerous biotechnological applications have been reported. Other notable advances include the generation of endonucleases and recombinases with altered specificities, made by innovative combinatorial and evolutionary protein engineering strategies. An unexpectedly high tolerance to mutation in the active sites of DNA polymerases is being exploited to engineer polymerases to incorporate artificial nucleotides or to display other, nonnatural activities.

Addresses
*Biochemistry and Molecular Biophysics, California Institute of Technology, mail code 210-41, Pasadena, California 91125, USA
†Division of Chemistry and Chemical Engineering, California Institute of Technology, mail code 210-41, Pasadena, California 91125, USA
‡e-mail: frances@cheme.caltech.edu

Introduction
Proteins involved in DNA recognition, manipulation and synthesis could provide a multitude of tools for studying gene function, genetic engineering, molecular biology and gene therapy. Engineered versions of these proteins would have the potential to modulate the expression of any gene of interest or to rearrange chromosomal DNA at any site within a genome. The utility of proteins that interact with DNA has been limited in many cases by their target sequence specificities. Thus, significant efforts have been made to engineer them to target alternate DNA sequences. DNA polymerases with altered fidelity and the ability to incorporate modified bases would enable new technologies for gene amplification, mutagenesis, and specific labeling. By engineering proteins that bind and modify DNA, we can also gain insights into the molecular mechanisms of maintenance, control and modification of genetic information. This review covers recent efforts to change the binding and catalytic specificities of transcription factors, endonucleases, recombinases, and DNA polymerases.

α-Helical DNA-binding proteins
The first report of engineering protein–DNA interactions involved modification of the helix–turn–helix DNA-binding domain of the 434 repressor so that it recognized the P22 operator [1]. Replacing the amino acids along the face of the recognition helix that make contacts with DNA with those from the corresponding positions in the P22 homolog created an engineered 434 repressor that bound the P22 operator with affinity similar to that of the wild-type P22 repressor for its cognate operator sequence. The engineered repressor no longer bound the 434 operator. Furthermore, heterodimers formed upon coexpression of the wild-type and modified 434 repressors could bind hybrid operators of 434 and P22 operator half-sites [2]. Single-chain proteins that are covalent dimers of the DNA-binding domains from different transcription factors have also been shown to bind to hybrid operators [3–9]. More recently, Liang et al. [10] constructed several single-chain heterodimers using engineered 434 repressor DNA-binding domains (identified in previous rational design and selection experiments), which also bound operators made up of half-sites bound by the individual domains. To achieve high DNA-binding affinities (Kd = 10−11–10−10 M), the single-chain proteins developed by Liang and colleagues require that the six base pair (bp) sequence between the two half-sites contains mostly A and T nucleotides. Apart from this study, there is little recent published work on engineering helix–turn–helix DNA-binding proteins. Thus, we cannot draw any conclusions about the ease of changing their binding specificities.

Zinc-finger proteins
One of the most abundant protein motifs in eukaryotes, the zinc-finger domain, has been the favored scaffold for engineering novel DNA-binding proteins [11,12]. A Cys2–His2 zinc finger consists of ~30 amino acid residues in a ββz fold stabilized by the coordination of two cysteine and two histidine residues to a zinc ion. Although each zinc finger typically recognizes only three DNA bases, multiple fingers can be linked in tandem so that the resulting multifinger protein can recognize longer

Abbreviations
bp base pair
CSR compartmentalized self-replication
dNTP deoxyribonucleotide triphosphate
PCR polymerase chain reaction
pol I polymerase I
rNTP ribonucleotide triphosphate
ZFP zinc-finger protein
sequences. This modularity is attractive because it opens the possibility of generating DNA-binding proteins of arbitrary sequence specificities by fusing pre-made fingers that each recognize any one of the 64 possible DNA triplets. Although this simplest vision of combinatorial zinc-finger protein (ZFP) design has only been partially realized, significant advances in design strategies have enabled the construction of artificial polydactyl ZFPs with diverse sequence specificities.

Phage display has been by far the most widely used technique to design (or discover) ZFPs with novel sequence specificities. Typically, a set of amino acids that contact the DNA in one finger of the multifinger protein is randomly mutated, and the mutant fingers are selected for binding to a desired DNA triplet, often with negative selection against binding to other triplets. This approach identified a set of zinc-finger domains that can recognize the 16 possible combinations of 5'-GNN-3' [13]. Extending this approach to zinc fingers that recognize the remaining 48 triplets, however, has been complicated by cross-subsite interactions in which DNA bases are contacted by amino acid residues from neighboring fingers. Dreier et al. [14] extended the repertoire of custom, interchangeable fingers to recognize 5'-ANN-3' triplets by eliminating a cross-subsite contact from the finger fused to the C terminus. Other recent refinements of this approach [15,16] further facilitate the ZFP design process.

An alternative strategy for selecting artificial ZFPs using phage display was recently described by Isalan and colleagues [17**] who constructed two libraries of three-finger proteins based on transcription factor Zif268. Each library contained half the wild-type Zif268 sequence (one and a half fingers); the remaining half harbored randomized amino acids that contact five of the nine bases in the recognition sequence. The libraries were selected in parallel for binding to double-stranded DNA the sequence of which contained four bases recognized by unmodified Zif268 and five bases of the target sequence. The selected half-libraries were recombined in vitro and further selected for binding to the full target DNA sequence. This approach allows the selection of high-affinity three-finger domains optimized for cross-subsite interactions. It also allows high-throughput selection of multiple ZFPs, because the half-libraries used for the initial selection can be used universally. The authors report that the entire selection process takes approximately two weeks and is amenable to automation.

Although overshadowed by the intense activity in phage display based selection of ZFPs, notable progress using a rational design strategy was recently reported by Sera and Uranga [18]. A nondegenerate recognition code table that assigns specific amino acids at positions -1, 2, 3, and 6 (relative to the start of the recognition helix) of a zinc finger to arbitrary 4 bp sequences was devised. The antisense base of the fourth base pair is contacted by the amino acid in position 2 of the first finger, whereas the sense base (which is also the first base of the second, overlapping 4 bp unit) is contacted by position 6 of the second finger. Proposing that one can design artificial ZFPs for arbitrary target sequences using the universal table, Sera and Uranga tested ten three-finger proteins targeted to different 10 bp sequences. Five exhibited nanomolar affinities towards the desired sequences, and the functional constructs were reported to discriminate single base-pair changes. The results suggest a preference for GC-rich sequences. With further refinement, this approach may complement the existing combinatorial design strategies, which still require substantial labor to construct ZFPs with novel sequence specificities.

With these advances in design and the rapid accumulation of knowledge regarding their DNA-binding properties, custom-designed ZFPs are now finding applications as artificial transcription regulators. Genes of interest can be activated or repressed in cells transfected with ZFPs fused to appropriate effector domains. Recent progress includes controlling gene expression in plants [19–22], inhibiting virus replication by targeting critical regulatory processes [23,24], and activating a gene involved in angiogenesis in a mouse model [25]. Blancafort and colleagues [26**] recently described the large-scale screening of cells transfected with ZFP transcription activator libraries for various phenotypic markers, demonstrating a promising new tool for functional genomics. These efforts have elucidated some important criteria for successful in vivo applications of ZFP-based transcription factors. It is critically important, for example, that the target DNA sequence be within the chromatin-accessible region. Liu et al. [27] clearly demonstrated this point in their work which identified accessible regions of the genomic DNA (which may differ among cell types) using a DNase I hypersensitivity assay. It also appears that ZFPs with six fingers function better than those with three fingers in most cases, most likely due to stronger binding and slower degradation.

**Restriction enzymes**

Restriction endonucleases are indispensable in today’s molecular biology. Years of screening various microbial sources have yielded hundreds of restriction enzymes that are capable of recognizing specific DNA sequences four to eight bases long and which cleave phosphodiester bonds within or adjacent to the recognition site. Engineered restriction enzymes that recognize altered or expanded sequences are needed for applications in biotechnology and medicine. However, modifying the sequence specificities of restriction enzymes has proved challenging. For example, a recent attempt by Lanio and coworkers [28] to rationally expand the recognition sequence of EcoRV based on available structural data...
yielded variants with altered selectivity, but not the predicted one. This group had previously shown that directed evolution could yield variants that prefer AT-rich flanking sites over GC-flanked sites [29]. The formidable challenges of rational engineering have led several researchers to turn to combinatorial or evolutionary methods, involving generation of random or directed mutant libraries coupled with screening or selection. However, because most restriction and other DNA-modifying enzymes do not exhibit the modular separation of functions that, for example, the ZFP-based transcription factors do, laboratory evolution of these enzymes must face the serious challenge of retaining catalytic activity while changing substrate sequence specificity. Phage display is therefore not useful, unless it can also select for catalysis [30*].

Samuelson and Xu [31] narrowed the substrate specificity of the promiscuous restriction enzyme BsaYI that cleaves four DNA sequences 5′- (A/G)GATC(C/T)-3′ with similar efficiency. BsaYI, N4-cytosine methyltransferase was used to protect 5′-AGATCT-3′ sites in the genome of host bacteria transformed with random mutants of BsaYI. Cells containing BsaYI mutants that retain activity toward the other three substrate sequences do not survive, due to damage to the genomic DNA. Mutant enzymes in the surviving cells, however, may or may not have retained activity toward 5′-AGATCT-3′ sites, which were protected by methylation. The mutants that survived the initial selection were subsequently tested for their ability to cleave the targeted 5′-AGATCT-3′ sequence using in vivo and in vitro screens. With further recombination and analysis of some functional mutations, a variant with at least 12-fold greater catalytic efficiency towards the targeted sequence was found. This particular variant, however, lost a significant fraction of its specific activity.

Random mutagenesis coupled with well-designed genetic assays allowed Seligman et al. [32] to identify several mutations that altered the target sequence specificity of homing endonuclease I-CreI. Homing endonucleases are encoded in introns or expressed as inteins of certain genes and are involved in the lateral gene transfer of their own genetic elements to the alleles that lack the intervening sequences [33]. The researchers incorporated the 22 bp homing site and mutated analogs into F′ plasmids containing kanamycin resistance and lacZ genes. I-CreI mutants that cleaved the wild-type homing sequence or its analogs would lose kanamycin resistance or yield white colonies in media containing 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-gal). Screening of mostly single amino acid mutants of I-CreI at positions known to make direct contacts with DNA bases identified several with altered or relaxed sequence specificities.

A more drastic modification of restriction sequence specificity was achieved by swapping domains of two homing endonucleases that naturally function as homodimers. Chevalier and colleagues [34**] took advantage of the fact that endonucleases of the LAGLIDADG family share a characteristic dimerization interface mediated by an α-helical pair to construct a hybrid enzyme stabilized by this common interface. An artificial single-chain enzyme that recognizes and cleaves hybrid DNA sequences was made by fusing the N-terminal domain of I-DmsoI to a I-CreI monomer via a short peptide linker (Figure 1). The hydrophobic interface of the chimeric protein, E-DrelI, was optimized for packing using an automated computational algorithm. Sixteen candidates were constructed and screened for solubility in vivo, and three were overexpressed and purified for further biochemical analysis in vitro. All three specifically recognized and cleaved 23 bp hybrid DNA sequences derived from the target sequences of the parental enzymes. The single turnover rate (kcat) of one chimera was comparable with that of the wild-type I-CreI, but the dissociation constant (Kd) towards the target DNA was two orders of magnitude higher than that of I-CreI towards its natural substrate sequence.

**Recombinases**

Recombinases integrate, excise, invert or translocate DNA based on the relative location and orientation of a target DNA site, which is typically palindromic and has a short spacer between half-sites. Used for site-specific recombination of DNA in prokaryotes, yeast, flies and mammals [35–37], their applications are currently limited by their strict target-site specificities. Evolutionary methods have been used with considerable success to engineer recombinases to target alternative DNA sites. Most screening methods require the mutant recombinase to excise a region of DNA that either places a reporter gene downstream of a constitutive promoter or removes a reporter gene that would be expressed in the absence of recombinase activity. Changes in the expression of the reporter gene are used to assess whether or not the mutant recombinase is functional on the desired DNA target site [38**,39*,40*,41]. The usual outcome of such experiments is relaxation of the target-site specificity, unless additional screening identifies those mutants that no longer function at the wild-type recombination site. Santoro and Schultz [40*] devised a method that allowed them to screen for Cre mutants that recombined at an engineered target site or that no longer recombined at the wild-type target site. When they alternated screening for recognition of the engineered site with screening for the inability to recognize the wild-type site in subsequent generations, they identified mutants with shifted DNA sequence specificities. When they did not specifically screen for loss of wild-type function, they always found mutants that were able to recombine at both the engineered and wild-type DNA sites. Towards the same goal of shifting rather than relaxing recombinase specificity, Buchholz and Stewart...
Computational interface redesign

In vitro biochemical characterization

In vivo solubility assay

In vitro biochemical characterization

Crystal structure determination and analysis

E-Drel (Engineered I-Dmol/I-Crel)
[42] placed the wild-type (loxP) and engineered (loxH) recombination sites on a single plasmid spaced in such a way that only one site could be recombined. Their competition-based approach generated Cre mutants with higher specificity for the engineered sequence. The strategy used in the most recent work on Flp recombinase includes a dual reporter system [38**,39*], which allows both wild-type and mutant-site recombination to be assessed simultaneously. This screen is based on the removal of two reporter genes by the recombinase. One reporter plasmid contains mutant recombination sites flanking the $lacZ_a$ gene, while a second reporter contains the wild-type recombination sites flanking the gene for red fluorescent protein (RFP). The authors identified Flp mutants that were active only at the mutant sites by selecting colonies expressing RFP but not $\beta$-galactosidase.

Several groups have used two different stepwise approaches to alter substrate specificity. In one approach, the recognition of the palindromic repeats is changed first followed by that of the directional spacer sequence [42]. In the second approach, mutants are progressively required to act at a target site with more mutations [38**]. Such stepwise strategies are likely to prove necessary as recombinases are engineered to recognize more drastically altered target sequences. Interestingly, Vozianov et al. [38**] found that the amino acid substitutions found in single target site mutants could not be recombined to produce a Flp variant that was active on a target site containing two mutations. Recombination with wild-type Flp and another round of random mutagenesis and screening, however, identified Flp variants able to recombine at the target sequence containing both mutations. One mutant showed a clear preference for this site over the wild-type and single-mutant sites.

Most efforts to change DNA sequence specificity have focused on mutating amino acid residues that make direct contact with the DNA. The identification of several non-contact positions that play key roles in determining the substrate specificity of Cre and Flp, however, indicates that this may not be the best strategy [38**,41]. The mutations shown to alter DNA target specificity of Flp (Figure 2) are clearly not limited to those that make direct contact with the DNA. Screening or selecting random mutagenesis libraries may be a useful addition to future efforts to engineer DNA sequence specificity.

DNA polymerases

DNA polymerases are integral to many molecular biology techniques, including sequencing, labeling, modification, amplification, detection, and random mutagenesis of targeted DNA. Potential uses for DNA polymerases also include the synthesis of DNA-based or DNA-like polymeric materials. Needs for higher-performance and
specialized polymerases have driven polymerase engineering efforts focused on altering properties such as processivity, activity, stability, and fidelity. Recent experiments have overwhelmingly used combinatorial or evolutionary approaches. Screens or selection systems for polymerase engineering, like those used for engineering recombine and endonuclease activities, must be based on enzyme function and not just on substrate binding.

Loeb and colleagues [43**] have conducted systematic mutational analyses of Escherichia coli and Taq polymerase I (pol I) to elucidate the molecular basis of their replication fidelity and substrate tolerance. Their experiments couple intensive mutagenesis of amino acid residues that contact the incoming deoxynucleotide triphosphates (dNTPs) with functional complementation of a polymerase-mutant E. coli [43**] or yeast strain [44]. Sequence analysis of some of the thousands of polymerase variants that are catalytically active has shown that several highly conserved residues are, nonetheless, tolerant to mutation. Hundreds of active mutants have been individually characterized to determine their activity, fidelity, and substrate specificity. Various interesting polymerases, such as error-prone E. coli pol I variants [45] and a Taq pol I that preferentially incorporates ribonucleotide triphosphates (rNTPs; by approximately 1000-fold) [46] have been identified. This approach, of intensive mutagenesis and genetic selection with subsequent biochemical analysis, was also used to analyze mouse polymerase β and resulted in the discovery of a variant with approximately 25-fold increased catalytic activity [47]. Although the active sites of DNA polymerases are strictly conserved in nature, where the selective pressure is apparently much more stringent than the laboratory genetic complementation, mutations in these sites can alter polymerase properties without destroying catalytic ability.

Polymerases involved in DNA repair are attractive for their broad substrate tolerance. Polymerase η is known for efficiently bypassing bulky adducts, such as cis-syn thymine dimers, and is one of the most error-prone of the polymerases (~10% error frequency). Glick et al. [44] developed a genetic selection, where inactive mutants rescue a UV-sensitive yeast strain deficient in its DNA repair system, to find functional polymerase η variants. From the functional variants, the authors isolated one with fourfold improved activity [44] and several better able to incorporate fluorescent dNTP analogs [48]. They also isolated mutants with 15-fold higher replication fidelity [49]. At present, these error-prone polymerases are not very practical for biotechnology applications due to their extremely slow polymerase activity (~10,000 times slower than typical pol I). However, the unique abilities of these enzymes to bypass damaged or irregular sites have proven useful in combination with other polymerases [50].

The fact that DNA polymerases can amplify their own genes establishes a link between genotype and phenotype, which can be used in evolutionary engineering to identify polymerases that are better self-replicators. Holliger and coworkers [51**] used the technique of compartmentalized self-replication (CSR). Polymerase variants are generated by in vitro mutagenesis and transformation into E. coli cells. The mutant polymerases are then encapsulated individually in droplets in a water/oil emulsion. In the droplet compartments, which also contain the components required for the polymerase chain reaction (PCR), the polymerases amplify only their own genes; here, improved function directly translates into gene amplification. Three cycles of CSR, in which the compartmentalized library pool was treated for a progressively longer time at 99°C (up to 15 min) before PCR, generated a Taq polymerase I with a half-life 11-fold greater than that of wild-type at 97.5°C. CSR also identified a Taq pol I variant with >130-fold increased resistance to heparin, a general DNA polymerase inhibitor.

A novel application of phage display allowed Romesburg and colleagues [30*] to select for polymerases that bind to rNTPs and act as RNA polymerases. They created a mutant library of Taq DNA pol I by fusion to the phage pIII coat protein. The substrate DNA template/primer duplexes were attached to other, adjacent pIII coat proteins. Polymerases that could extend the attached oligonucleotide primer by incorporating rNTP and biotinylated rUTP were selectively recovered using streptavidin-coated magnetic beads. Four rounds of screening isolated mutants that incorporate rNTPs virtually as efficiently as the wild-type enzyme incorporates dNTP substrates. Notably, each rNTP was incorporated with similar efficiency, although the wild-type Taq pol I and variants reported elsewhere [46] show very poor incorporation of rUTP. None of these polymerases, however, show good processivity.

DNA replication, the raison d’etre of DNA polymerases, is an extraordinarily accurate process. The rate of somatic mutation in mammalian cells is estimated to be about 10^{-10} per cell per duplication, and even very subtle changes in polymerase properties could lead to genetic instability. Nonetheless, the studies discussed here show that polymerases are quite robust to mutation. While the evolutionary implications of this fact remain elusive, these enzymes have significant engineering potential.

Conclusions

Engineering the target specificities of proteins that bind and modify DNA has proven challenging, but by no means impossible. Innovative structure-based and evolutionary design strategies have generated new transcription factors, restriction enzymes, recombinases and polymerases, and we can begin to envision engineering proteins capable of binding and acting at any (accessible) target DNA sequence. With these proteins will come the
power to control the expression of any gene or to recombine, excise or incorporate new DNA at specified sites within a genome. Engineered DNA polymerases will lead to improvements in such essential processes as PCR and mutant library generation, as well as fundamentally new applications. It is clear that we are just beginning to examine and understand how ‘designable’ these proteins are and to identify the most effective methods for engineering them. Current work, however, indicates a level of functional plasticity which, coupled with the multitude of potential applications, promises significant advances in the next few years.

Acknowledgements

We acknowledge support from the US National Science Foundation (Biological Information Technology and Storage (BITS) Grant EIA-0130613) and the US Defence Advanced Research Project Agency (DARPA) under award N66001-02-1-8929 (Disclaimer: any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the DARPA). YY also acknowledges support from a Japan Society for the Promotion of Science Postdoctoral Fellowship for Research Abroad.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest


This paper describes a clever phage display approach to select for RNA polymerase activity.


34. Chevalier BS, Kortemme T, Chaday MS, Baker D, Monnat RJ. ** Stoddard BL: ** Design, activity, and structure of a highly specific artificial endonuclease. Mol Cell 2002, 10:895-905. An impressive protein engineering effort to construct chimeric endonucleases with novel, predicted specificities using an automated computational algorithm. The crystal structure of one of the chimeric enzymes supports the design process.


38. Voziyanov Y, Konieczka JH, Stewart AF, Jayaram M: Stepwise manipulation of DNA specificity in Flp: progressively adapting Flp to individual and combinatorial mutations in its target site. J Mol Biol 2003, 326:65-76. This work combines rational and evolutionary design strategies to engineer the DNA target-site specificity of Flp recombinase. A powerful double-reporter screen was used to identify mutants with altered substrate specificities.

39. Voziyanov Y, Stewart AF, Jayaram M: A dual reporter screening system identifies the amino acid at position 82 in Flp site-specific recombinase as a determinant for target specificity. Nucleic Acids Res 2002, 30:1666-1663. This paper describes a dual reporter screen for identifying Flp recombinase variants that recognize altered target DNA sequences. One reporter gene is expressed constitutively until it is excised by a recombinase that recognizes the wild-type target sequence. A second reporter gene is expressed constitutively until it is excised by a recombinase that recognizes a novel target DNA sequence. This system was used to identify mutants with either relaxed or shifted substrate specificity. By comparing these proteins, the authors identified an amino acid that plays an important role in determining substrate specificity.

40. Santoro SW, Schultz PG: Directed evolution of the site specificity of Cre recombinase. Proc Natl Acad Sci USA 2002, 99:4185-4190. The authors used a directed evolution strategy to identify Cre recombinase variants that recognize altered DNA target sites. In separate screens, they identified Cre mutants that recognized the novel target sites (positive screening) and variants that could not recognize the wild-type site (negative screening). A combination of positive and negative screening was required to switch specificity.


51. Ghadessy FJ, Ong JL, Holliger P: Directed evolution of polymerase function by compartmentalized self-replication. Proc Natl Acad Sci USA 2001, 98:4552-4557. This paper demonstrates the use of CSR for evolving polymerase function. With a heat-stable water-in-oil emulsion system to encapsulate mutant polymerases with PCR reagents, each DNA polymerase can be amplified according to its activity. This system was used to evolve polymerases with higher thermostability or resistance to an inhibitor.
Erratum
Engineering proteins that bind, move, make and break DNA
Cynthia H Collins, Yohei Yokobayashi, Daisuke Umeno and Frances H Arnold

In this article, published in the August 2003 issue of Current Opinion in Biotechnology, the author’s discussion of engineering α-helical DNA-binding domains failed to note earlier reports of engineering protein–DNA interactions [1–3]. In particular, two studies reported altering the DNA-binding specificities of the Mnt repressor [1] and catabolite activator protein [2], using genetic selection methods, whereas a third described the substitution of an entire α-helix to alter the DNA-binding specificity of the 434 repressor protein [3]. The authors regret this significant oversight.

References