# Designing Extensible Protein-DNA Interactions for Synthetic Biology

Kristjan E. Kaseniit, Samuel D. Perli, Timothy K. Lu*
Synthetic Biology Group, Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA USA
timlu@mit.edu

***Abstract*—Protein-DNA interactions are essential to constructing biological devices for synthetic gene circuits. Ideal devices should be interoperable and extensible with respect to each other. They should also exhibit minimal unwanted interactions with the host cells in which they reside and be portable between different host chassis. Here, we discuss two classes of protein-DNA devices, memory modules and transcription factors, that can be used to construct genetic circuits with novel functionalities. In addition, we describe a methodology for identifying candidate DNA targets to be used in designing artificial protein-DNA interactions. We identified 9 base-pair (bp) sites within each of six useful host organisms that were absent in individual host genomes but were unable to find any 9 bp sites that were absent from all of the genomes. Extending our search to 12 bp and 15 bp DNA sequences revealed tens of thousands and millions, respectively, of DNA sequences that were absent in all host organisms we evaluated; these sequences were targetable by publicly accessible zinc-finger methodologies. By targeting these sites, it may be possible to build interoperable, orthogonal, and portable protein-DNA devices. This work lays a foundation for future efforts to engineer novel biological devices in the emerging field of synthetic biology.**

## I. Introduction

Engineering biological circuits differs from electrical circuits in several significant ways [1, 2]. First, biological "wiring" is conferred by chemical specificity and localization rather than physical connections as in electronics. Second, electrical parts can be rapidly instantiated while well-characterized biological parts are relatively scarce. Third, the rules that govern fundamental concepts such as modularity, orthogonality, and reliability are not well understood in biological systems compared with electronics.

Protein-DNA interactions are crucial in determining the specificity of many biological parts, including recombinases and transcription factors. Since biological circuits operate in the context of their host cellular chassis, one must take potential interactions with cellular proteins and DNA, in addition to other synthetic devices, into account when designing biological systems. In order to design synthetic parts which are less likely to interact with the host chassis, we have identified DNA sequences which are absent or under-represented in host genomes. Ideally, we would like to target DNA sequences which are absent across multiple host species to maximize the potential portability of the resulting devices.

## II. Single-Invertase Memory Modules (SIMMs)

We have previously designed Single Invertase Memory Modules (SIMMs) to encode memory into the DNA of living cells (Figure 1) [3]. A SIMM is composed of a gene encoding a bidirectional recombinase, of which >100 are known [4], in between two copies of the recombinase's DNA recognition sites. The DNA sites are oriented towards each other such that

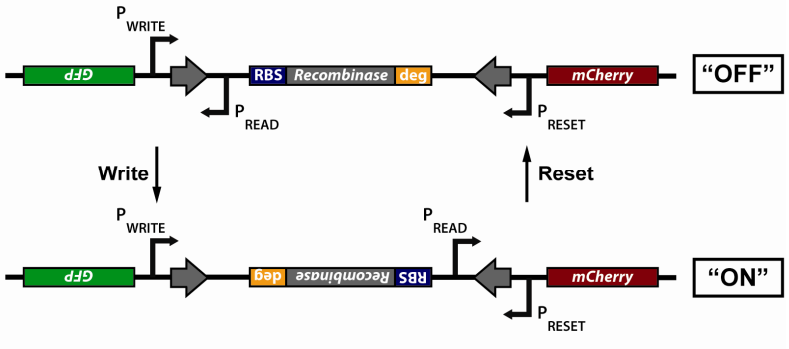


Figure 1. The Single Invertase Memory Module (SIMM) has opposing recombinase recognition sites (grey arrows) which contain between them an inverted promoter ($P_{READ}$), a synthetic ribosome-binding-sequence (RBS), a recombinase gene (*Recombinase*) and an ssrA-based degradation tag (deg). A SIMM maintains memory based on DNA orientation, which is inverted when the recombinase is expressed, enabling two distinct states. Upon flipping, the originally inverted promoter ($P_{READ}$) becomes upright, enabling the expression of downstream genes as outputs. A two-color system, as shown here, can be used to characterize SIMMs. The $P_{READ}$ promoters produce GFP or mCherry fluorescence depending on SIMM orientation. The RBS and degradation tags control the expression and temporal stability of the bidirectional recombinases. This design is extensible due to >100 known recombinases.

This work was supported in part by the MRSEC Program of the National Science Foundation under award number DMR - 0819762. This work was also supported in part by the Office of Naval Research.

recombinase activity achieves inversion rather than excision due to the fundamental biochemical mechanisms of recombinases proteins. Upon expression of the recombinase by an upstream promoter, the SIMM inverts and further recombinase expression ceases. This biological memory architecture resembles static RAM (SRAM) in that it only requires active transcription and translation to transition between memory states but not for the maintenance of memory, which is hardcoded into DNA orientation. This is in contrast to other biological memory designs, such as the toggle switch (Figure 2) [5], which are analogous to dynamic RAM (DRAM) in that they require active transcription and translation to maintain their epigenetic state. In addition, each bit of memory encoded by a toggle switch requires two transcriptional repressor proteins, while each bit of memory encoded by a SIMM requires only one recombinase (Figure 2). This is an advantage in favor of SIMMs due to the shortage of parts in synthetic biology.

SIMMs achieve interoperability and extensibility because each recombinase protein type has unique DNA target sites which are not recognized by other recombinase protein types. Thus, multiple recombinases can be used together in a single cell without significant interference. To extend the library of available recombinases, investigators have explored the design of artificial recombinases. Artificial site-specific recombinases which act upon specific DNA sites can be constructed by combining DNA-targeting domains with recombinase effector domains [6-8]. Such DNA-targeting domains must be engineered to be orthogonal with each other and the host chassis in order to enable proper operation.

## III. Synthetic Transcription Factors

In addition to recombinase-based memory modules, transcription factors constitute a major class of biological devices. Transcription factors rely on targeted protein-DNA interactions in order to maintain specificity and avoid crosstalk. Transcription factors can be built by fusing DNA-binding domains and transcriptional effector domains together into a single protein. For example, Joung *et al.* described a bacterial two-hybrid (B2H) system in which zinc-finger (ZF) arrays bind to specific DNA targets upstream of a weak promoter and recruit RNA polymerase to activate transcription [9]. Others, including us, have designed zinc-finger and TAL-effector transcription factors in eukaryotes [10-16].

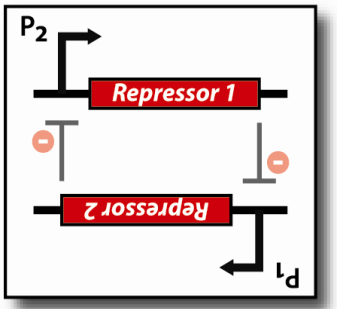


Figure 2. The toggle switch involves two transcriptional repressors, each of which suppresses the expression of the other. This architecture can produce bistable states that can be toggled by the addition of inducer molecules that deactivate one of the repressors. It requires active transcription and translation to maintain memory.

$Cys_2$-$His_2$ zinc fingers (ZFs) are useful elements for implementing orthogonal and tunable regulatory elements for synthetic biology. The DNA-binding specificities of ZFs can be engineered to target a wide range of sequences. Moreover, they can be used to target longer DNA sequences by covalently linking them into multi-finger arrays [10, 17-19]. Several robust and public platforms are available for engineering arrays with multiple zinc finger, including the Oligomerized Pool Engineering (OPEN) [20, 21] and Context-Dependent Assembly (CoDA) [19] methods. Proprietary libraries of zinc-finger arrays have also been described and used for a variety of genome-targeting applications [22, 23].

In contrast to previously described selection-based approaches [24, 25], OPEN obviates the need to construct and test very large combinatorial ZF libraries (>$10^8$ in size). This is a significant advantage that enables OPEN to performed very rapidly while maintaining high-quality ZF arrays. OPEN utilizes a library of pre-validated ZF pools (consisting of ZFs which are targeted to a given three base-pair "subsite" at a defined position within a three-finger protein). These pools are then shuffled to create a small but diverse library of variant ZF arrays. This library is probed with a B2H selection system where the binding of a ZF domain to it a target DNA site activates the expression of selectable marker genes [9]. Following this round of selection, a quantitative *lacZ* reporter system is used to measure how well each resulting ZF array activates transcription [26].

Recently, the Joung lab also described CoDA [19], which is a simplified platform of publicly available reagents and software for constructing ZF arrays. The success rate of CoDA is comparable to selection-based methods while being rapid and requiring no specialized expertise. With standard molecular cloning techniques or DNA synthesis, ZF arrays can be constructed with CoDA in one to two weeks or less. The CoDA system has been validated as a robust ZF engineering platform by being used to successfully engineer arrays for over 200 different 9 base-pair target sites.

In addition to ZFs, other modular protein domains can be used for programmable DNA binding. For instance, TAL (transcription activator-like) domains from *Xanthomonas* bacteria can be engineered readily [16]. TAL domains contain ~34 amino-acid residues. Each domain binds to a single base-pair of DNA; the specificity of binding is determined by the identities of two amino acids [27, 28]. Like ZFs, more extended arrays capable of binding to longer DNA sequences can be constructed by linking multiple TAL domains together; this capability has been validated by engineering TAL arrays fused to nucleases for inducing targeted alterations in endogenous human genes [27, 29-31].

## IV. Computational Search for Under-Represented DNA Targets

Synthetic biological devices, such as recombinases and transcription factors, should exhibit minimal undesirable interactions with the host chassis in which they reside. Thus, we sought to assemble a library of potential DNA targets with the hypothesis that these DNA targets should be absent or under-represented in the genomes of host organisms to decrease the probability of unwanted device-host interactions. This approach for constructing synthetic protein-DNA interactions is substantially different than ones described for targeting endogenous DNA sites.

To isolate under-represented DNA sites, we used available FASTA files for the hosts *Escherichia coli* str. K-12 substr. MG1655, *Bacillus subtilis* BSn5, *Saccharomyces cerevisiae* S288c, *Pichia pastoris* GS115, *Schizosaccharomyces pombe* 972h and *Kluyveromyces lactis* NRRL Y-1140. These hosts were chosen due to their potential utility for biotechnology and synthetic biology. From these genomes, we extracted the number of instances of all the 9 base-pair (bp) and 12 bp DNA sequences possible.

The program used for this approach sought to determine the number of occurrences of a given physical double-stranded DNA segment in a genome. This was done by using a sliding-window with a fixed size to represent one strand of DNA and counting how many times the strand or its complement (ordered 5' to 3') appeared in a given FASTA file. Care was taken to count combinations arising from the circular property of some of the DNA.

All sequences found in the genomes that contained IUPAC nucleotides other than A, T, G and C (e.g. N, M or R) were not added to the database. The final database took into account all the possible 9 bp and 12 bp double-stranded sequences per chromosome or mitochondrion in each organism. Thus, the databases effectively covered all possible 262,144 and 16,777,216 combinations, respectively, while accounting for the fact that a given sequence represents the same physical double-stranded DNA as its reverse complement.

From this database we could perform multiple queries, such as finding 9 bp DNA target sites which occur in genomes only 0 to 2 times. For each candidate site, we also evaluated whether available subsites or their complements were available for targeting by zinc-finger arrays with the OPEN method by referencing the ZiFiT zinc-finger array designer software [32]. The ZiFDB database was also used to identify candidate sites for which existing zinc-finger arrays have been described [33]. We found no 9 bp sites that were absent from all of the organisms tested, suggesting that 9 bp may be too short of a sequence with which to develop synthetic protein-DNA interactions which are portable and orthogonal with respect to host genomes. Table I summarizes this data.

TABLE I. Number of 9 bp DNA Sites in Host Organisms

| Organism | # of instances in genome | # of such 9 bp sequences | Of those, # in ZiFiT/ZiFDB |
|---|---|---|---|
| *E. coli* | 0 | 1251 | 114 |
| *B. subtilis* | 0 | 299 | 27 |
| *S. cerevisiae* | 0 | 1 | 0 |
| *P. pastoris* | 0 | 15 | 2 |
| *S. pombe* | 0 | 38 | 7 |
| *K. lactis* | 0 | 53 | 3 |
| SUBTOTAL | 0 | 1657 | 153 |
| *E. coli* | 1 | 1482 | 101 |
| *B. subtilis* | 1 | 776 | 89 |
| *S. cerevisiae* | 1 | 8 | 0 |
| *P. pastoris* | 1 | 117 | 11 |
| *S. pombe* | 1 | 128 | 18 |
| *K. lactis* | 1 | 186 | 18 |
| SUBTOTAL | 1 | 2697 | 237 |
| *E. coli* | 2 | 1433 | 109 |
| *B. subtilis* | 2 | 1393 | 143 |
| *S. cerevisiae* | 2 | 26 | 1 |
| *P. pastoris* | 2 | 203 | 19 |
| *S. pombe* | 2 | 229 | 28 |
| *K. lactis* | 2 | 341 | 34 |
| SUBTOTAL | 2 | 3625 | 334 |
| TOTAL | 0-2 | 7979 | 724 |

For evaluating the availability of 12 bp candidate sites, we chose to look at 0-instance sequences in the genomes of the 6 organisms described above. A sequence was considered available for construction if the first 9 bp subsequence or the last 9 bp subsequence was available in ZiFiT's OPEN method and the rest of the sequence, a triplet, was available in ZiFiT's OPEN method for any position. This was done because the ZiFiT's OPEN method is currently focused on three-finger arrays and we assumed that a four-finger array to target 12 bp DNA sites could be constructed by fusing an additional zinc-finger to the array.

We found 25,684 available 12 bp candidate sites where the first 9 bp subsequence was in ZiFiT and 26,438 candidate sites where the last 9 bp subsequence was in ZiFiT. The total number of candidate sites was 35,632. The following is a sample of some of these candidate DNA sites:

5'GGCGCGTGT*TTT*3', 5'GCCGCGTGT*TTT*3', 5'GCTTAGGGG*TTT*3', 5'*TAG*GGGGCTTTT3', 5'*GGG*TCTGATTTT3', 5'*GGG*GCCTTGTTT3' where italicized base-pairs represent ones for which the ZiFiT assumption described above was made.

We also considered 15 bp sites which contained a viable 9 bp site with the rest of the sequence filled with triplets from the ZiFiT OPEN program. In addition to the 6 organisms mentioned above, we filtered out any sequences that appeared in *Mus musculus* or *Homo sapiens*. We found 3,111,653 0-instance 15 bp sites that were available for targeting by zinc-finger arrays constructed using OPEN.

## CONCLUSIONS

We have presented two device architectures, SIMMs and transcription factors, where programmable, portable, and orthogonal protein-DNA interactions are desirable. A computational search of six bacterial and yeast host organisms revealed that no 9 bp site that was absent from all of the host genomes could be found. Expanding the search to 12 bp and 15 bp sequences revealed tens of thousands to millions of potential target sites. These sites will form the foundation for our future efforts to engineer artificial protein-DNA interactions for synthetic biology parts.

## ACKNOWLEDGMENTS

We would like to thank members of the Lu lab for their comments and suggestions and the Thomas and Sarah Kailath Fellowship Fund for their support to Samuel D. Perli.

## REFERENCES

[1] T. K. Lu, "Engineering Scalable Biological Systems," *Bioengineered Bugs,* vol. 1, pp. 378-384, November/December 2010.

[2] T. K. Lu, *et al.*, "Next-generation synthetic gene networks," *Nat Biotechnol,* vol. 27, pp. 1139-50, Dec 2009.

[3] A. E. Friedland, *et al.*, "Synthetic gene networks that count," *Science,* vol. 324, pp. 1199-202, May 29 2009.

[4] A. C. Groth and M. P. Calos, "Phage integrases: biology and applications," *J Mol Biol,* vol. 335, pp. 667-78, Jan 16 2004.

[5] T. S. Gardner, *et al.*, "Construction of a genetic toggle switch in Escherichia coli," *Nature,* vol. 403, pp. 339-42, Jan 20 2000.

[6] C. Proudfoot, *et al.*, "Zinc finger recombinases with adaptable DNA sequence specificity," *PLoS One,* vol. 6, p. e19537, 2011.

[7] H. J. Lee, *et al.*, "Site-specific DNA excision via engineered zinc finger nucleases," *Trends Biotechnol,* vol. 28, pp. 445-6, Sep 2010.

[8] R. M. Gordley, *et al.*, "Synthesis of programmable integrases," *Proc Natl Acad Sci U S A,* vol. 106, pp. 5053-8, Mar 31 2009.

[9] J. K. Joung, *et al.*, "A bacterial two-hybrid selection system for studying protein–DNA and protein–protein interactions," *Proceedings of the National Academy of Sciences,* vol. 97, pp. 7382-7387, June 20, 2000.

[10] R. R. Beerli and C. F. Barbas, 3rd, "Engineering polydactyl zinc-finger transcription factors," *Nat Biotechnol,* vol. 20, pp. 135-41, Feb 2002.

[11] S. Stolzenburg, *et al.*, "Modulation of gene expression using zinc finger-based artificial transcription factors," *Methods Mol Biol,* vol. 649, pp. 117-32, 2010.

[12] S. Kim, *et al.*, "Construction of combinatorial libraries that encode zinc finger-based transcription factors," *Methods Mol Biol,* vol. 649, pp. 133-47, 2010.

[13] J. S. Kang and J. S. Kim, "Zinc finger proteins as designer transcription factors," *J Biol Chem,* vol. 275, pp. 8742-8, Mar 24 2000.

[14] P. Blancafort, *et al.*, "Genetic reprogramming of tumor cells by zinc finger transcription factors," *Proc Natl Acad Sci U S A,* vol. 102, pp. 11716-21, Aug 16 2005.

[15] F. Zhang, *et al.*, "Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription," *Nat Biotechnol,* vol. 29, pp. 149-53, Feb 2011.

[16] H. Scholze and J. Boch, "TAL effectors are remote controls for gene activation," *Curr Opin Microbiol,* vol. 14, pp. 47-53, Feb 2011.

[17] A. C. Jamieson, *et al.*, "Drug discovery with engineered zinc-finger proteins," *Nat Rev Drug Discov,* vol. 2, pp. 361-8, May 2003.

[18] C. O. Pabo, *et al.*, "Design and selection of novel Cys2His2 zinc finger proteins," *Annu Rev Biochem,* vol. 70, pp. 313-40, 2001.

[19] J. D. Sander, *et al.*, "Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)," *Nat Methods,* vol. 8, pp. 67-9, Jan 2011.

[20] M. L. Maeder, *et al.*, "Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification," *Mol Cell,* vol. 31, pp. 294-301, Jul 25 2008.

[21] M. L. Maeder, *et al.*, "Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays," *Nat Protoc,* vol. 4, pp. 1471-501, 2009.

[22] C. L. Dent, *et al.*, "Regulation of endogenous gene expression using small molecule-controlled engineered zinc-finger protein transcription factors," *Gene Ther,* vol. 14, pp. 1362-1369, 2007.

[23] Y. Doyon, *et al.*, "Enhancing zinc-finger-nuclease activity with improved obligate heterodimeric architectures," *Nat Meth,* vol. 8, pp. 74-79, 2011.

[24] J. A. Hurt, *et al.*, "Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection," *Proc Natl Acad Sci U S A,* vol. 100, pp. 12271-6, Oct 14 2003.

[25] H. A. Greisman and C. O. Pabo, "A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites," *Science,* vol. 275, pp. 657-61, Jan 31 1997.

[26] D. A. Wright, *et al.*, "Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly," *Nat Protoc,* vol. 1, pp. 1637-52, 2006.

[27] J. Boch, *et al.*, "Breaking the code of DNA binding specificity of TAL-type III effectors," *Science,* vol. 326, pp. 1509-12, Dec 11 2009.

[28] M. J. Moscou and A. J. Bogdanove, "A simple cipher governs DNA recognition by TAL effectors," *Science,* vol. 326, p. 1501, Dec 11 2009.

[29] M. Christian, *et al.*, "Targeting DNA double-strand breaks with TAL effector nucleases," *Genetics,* vol. 186, pp. 757-61, Oct 2010.

[30] T. Li, *et al.*, "TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain," *Nucleic Acids Res,* vol. 39, pp. 359-72, Jan 1 2011.

[31] J. C. Miller, *et al.*, "A TALE nuclease architecture for efficient genome editing," *Nat Biotechnol,* Dec 22 2010.

[32] J. D. Sander, *et al.*, "ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool," *Nucleic Acids Research,* vol. 38, pp. W462-W468, July 1, 2010 2010.

[33] F. Fu, *et al.*, "Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays," *Nucleic Acids Research,* vol. 37, pp. D279-D283, January 1, 2009.