ACCOUNTS of chemical research

Article

Cite This: Acc. Chem. Res. 2018, 51, 394−403

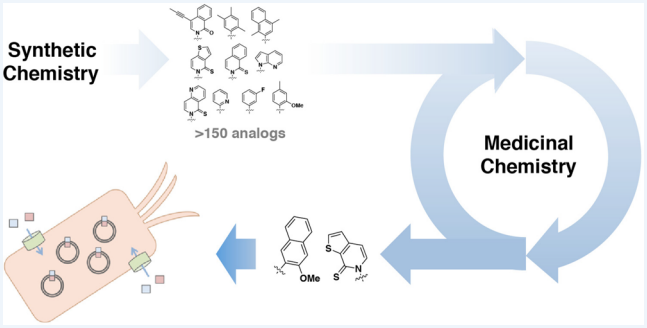
pubs.acs.org/accounts

Expansion of the Genetic Alphabet: A Chemist's Approach to Synthetic Biology

Aaron W. Feldman and Floyd E. Romesberg*

Department of Chemistry, The Scripps Research Institute, La Jolla, California 92037, United States

CONSPECTUS: The information available to any organism is encoded in a four nucleotide, two base pair genetic code. Since its earliest days, the field of synthetic biology has endeavored to impart organisms with novel attributes and functions, and perhaps the most fundamental approach to this goal is the creation of a fifth and sixth nucleotide that pair to form a third, unnatural base pair (UBP) and thus allow for the storage and retrieval of increased information. Achieving this goal, by definition, requires synthetic chemistry to create unnatural nucleotides and a medicinal chemistry-like approach to guide their optimization. With this perspective, almost 20 years ago we began designing unnatural nucleotides with the ultimate goal of developing UBPs that function *in vivo*, and thus serve as the foundation of semi-synthetic organisms (SSOs) capable of storing and retrieving increased information. From the beginning, our efforts focused on the development of nucleotides that bear predominantly hydrophobic nucleobases and thus that pair not based on the complementary hydrogen bonds that are so prominent among the natural base pairs but rather via hydrophobic and packing interactions. It was envisioned that such a pairing mechanism would provide a basal level of selectivity against pairing with natural nucleotides, which we expected would be the greatest challenge; however, this choice mandated starting with analogs that have little or no homology to their natural counterparts and that, perhaps not surprisingly, performed poorly. Progress toward their optimization was driven by the construction of structure−activity relationships, initially from *in vitro* steady-state kinetic analysis, then later from pre-steady-state and PCR-based assays, and ultimately from performance *in vivo*, with the results augmented three times with screens that explored combinations of the unnatural nucleotides that were too numerous to fully characterize individually. The structure−activity relationship data identified multiple features required by the UBP, and perhaps most prominent among them was a substituent ortho to the glycosidic linkage that is capable of both hydrophobic packing and hydrogen bonding, and nucleobases that stably stack with flanking natural nucleobases *in lieu* of the potentially more stabilizing stacking interactions afforded by cross strand intercalation. Most importantly, after the examination of hundreds of unnatural nucleotides and thousands of candidate UBPs, the efforts ultimately resulted in the identification of a family of UBPs that are well recognized by DNA polymerases when incorporated into DNA and that have been used to create SSOs that store and retrieve increased information. In addition to achieving a longstanding goal of synthetic biology, the results have important implications for our understanding of both the molecules and forces that can underlie biological processes, so long considered the purview of molecules benefiting from eons of evolution, and highlight the promise of applying the approaches and methodologies of synthetic and medical chemistry in the pursuit of synthetic biology.



INTRODUCTION

The field of synthetic biology was first defined in 1911 by Stéphane Leduc[1] with the goal of creating new biological forms and functions. The modern field is largely focused on using the engineering-like approach of design, test, and standardize to optimize naturally derived "parts", most commonly novel DNA elements. However, the most fundamental approach to create new forms and functions is to develop unnatural base pairs (UBPs) that expand the genetic alphabet and, thus, increase the amount of information that may be stored in an organism's DNA. The effort to expand the genetic alphabet was first championed by Steven Benner and focused on unnatural nucleotides bearing hydrogen-bonding (H-bonding) patterns that are orthogonal to those employed by the natural base pairs (Figure 1).[2] However, it was unclear if H-bonding was the only force suitable for controlling base pairing. Indeed, the Kool laboratory reported the remarkable observation that a DNA polymerase could selectively insert the difluorotoluene analog of dTTP opposite dA in the template.[3] We and the Hirao laboratory thus initiated efforts to use hydrophobic and packing forces to control UBP formation. Hirao has focused on derivatizing natural purine and pyrimidine scaffolds to create "shape complementary" UBPs (Figure 1),[4] while we focused on nucleotides bearing nucleobase analogs with little to no homology to their natural counterparts.

Although our efforts were consistent with the tenets of modern synthetic biology, we used synthetic chemistry to

Received: August 16, 2017
Published: December 2, 2017

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

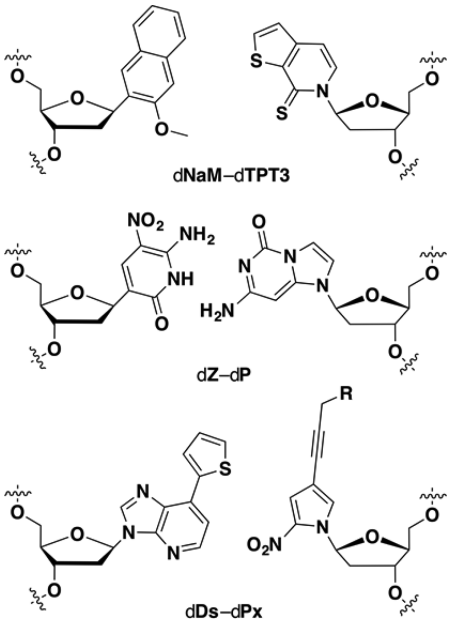


Figure 1. d**NaM**-d**TPT3**, d**Z**-d**P**, and d**Ds**-d**Px** (R = H or −CH(OH)−CH_2OH) UBPs.

generate the required parts, which is in many ways more consistent with Leduc's original vision.[1] Furthermore, we optimized the parts using a medicinal chemistry-like approach, inspired by the perspective that any effort endeavoring to develop foreign, man-made parts that function within a cell will need to optimize solubility, cellular uptake, stability, off-target activity, toxicity, dose, and dosing regimen. The field of medicinal chemistry approaches the same problems through the construction of structure−activity relationships (SARs) that allow for empirical optimization in the absence of a complete understanding of the process being optimized, a strategy that we adapted for UBP development.

Here, we recount our efforts to develop UBPs that ultimately culminated in the discovery of the d**NaM**−d**TPT3** UBP (Figure 1), as well as a family of related UBPs, all of which have now been used to create semi-synthetic organisms (SSOs) that store[5,6] and retrieve[7] increased information. This places us at the doorstep of realizing Leduc's vision of creating organisms with novel forms and functions.

PARTS OPTIMIZATION: FIRST GENERATION UBP CANDIDATES

The synthetic biology parts required for the expansion of the genetic alphabet are the unnatural nucleotides that selectively pair to form a UBP, and their optimization, at least initially, was measured by both duplex stability and DNA polymerase recognition. In most cases, analogs were synthesized as triphosphates (referred to as d**X**TP, where **X** is a nucleoside analog), as well as phosphoramidites for incorporation into DNA via solid phase synthesis. While characterization of the stability of duplex DNA containing the UBPs revealed many interesting trends, such as the effects of solvation,[8] stability proved uncorrelated with polymerase recognition, as is also the case with natural base pairs,[9] and polymerase recognition quickly emerged as our primary focus. The majority of our early efforts employed the exonuclease-deficient Klenow fragment of *E. coli* DNA polymerase I (Kf). We focused specifically on two steps, the insertion of an unnatural triphosphate opposite its cognate analog in a template (a step that we also refer to as UBP synthesis or unnatural nucleotide incorporation), and the continued primer extension by insertion of the next correct triphosphate, in each case characterizing efficiency (second-order rate constant, k_{cat}/K_M) and fidelity ($(k_{cat}/K_M)_{correct}/(k_{cat}/K_M)_{incorrect}$). Steady-state conditions were employed, which allowed rapid SAR construction but only provided information about the rate-limiting step of insertion or extension, both of which are actually a complex series of reactions (e.g., triphosphate binding, conformational changes, phosphoryl transfer, product release). During early development, this was not a problem as the rate-limiting step was invariably phosphoryl transfer.

Our search began with the simple benzene and naphthalene nucleobase analogs (Figure 2A).[10,11] We found that d**DM**TP is

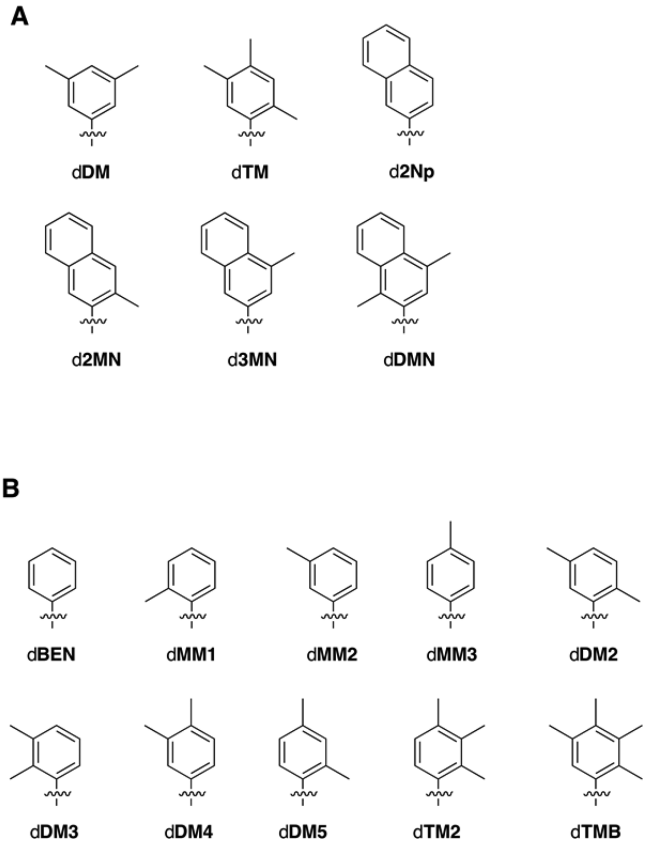


Figure 2. (A) First generation methylated benzene and naphthalene analogs. (B) Second generation methylated benzene analogs. Sugar and phosphate groups omitted for clarity.

a poor polymerase substrate, as incorporation opposite d**DM** or d**TM** in the template was virtually undetectable. In contrast, d**TM**TP was more efficiently incorporated opposite d**DM** ($k_{cat}/K_M = 1.4 \times 10^6$ M^{-1} min^{-1}) and d**TM** ($k_{cat}/K_M = 2.2 \times 10^6$ M^{-1} min^{-1}), only 20−30-fold less than the efficiency with which a natural base pair is synthesized in the same sequence context. However, both the d**TM**−d**DM** heteropair and d**TM**−d**TM** self-pair are limited by poor extension and mispairing with dA, likely because dA is the most hydrophobic of the natural nucleotides. We also found that d**2Np** efficiently directs the insertion of d**2Np**TP ($k_{cat}/K_M = 2.8 \times 10^6$ M^{-1} min^{-1}), but misincorporation of dATP was again problematic ($k_{cat}/K_M = 1.1 \times 10^5$ M^{-1} min^{-1}). Addition of a methyl substituent to the position ortho to the glycosidic bond (hereafter referred to

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

simply as the ortho position) yielded d**2MN**, which generally increased the rate of incorporation of the triphosphate against hydrophobic analogs in the template. However, d**2MN**TP is again efficiently inserted opposite dA. Moving the single methyl substituent from the ortho to the meta position (d**3MN**) reduces insertion opposite dA but also reduces the efficiency of UBP synthesis. The additional methyl substituent of d**DMN** resulted in efficient pairing with both d**TM** and d**2MN**, but it also increased the rate of dATP insertion when in the template. Thus, while mispairing generally remained problematic, at this point it was clear that several of these UBPs showed promising synthesis rates. However, none of them supported continued primer extension at a detectable level (k_{cat}/K_M < 10^3 M^{-1} min^{-1}).

The isocarbostyril scaffold also received early development attention (Figure 3A,B).[10,12−14] When in the template, the simplest of the series, d**ICS**, directs the insertion of triphosphates bearing simple substituted benzene nucleobases, such as d**DM**TP, d**TM**TP, or d**DMN**TP, with only modest efficiency, but d**2MN**TP is inserted significantly more efficiently. Addition of a 1-propynyl group to the 7-position of d**ICS**, affording d**PICS**, results in a triphosphate that is generally inserted more efficiently opposite other unnatural analogs. The methyl group of d**PIM** resulted in efficient but indiscriminate insertion of the triphosphate, while addition of the methyl group of d**MICS** resulted in the indiscriminate templating of natural triphosphates. Although the methyl group of d**5MICS** had little systematic effect on UBP synthesis, it did dramatically reduce self-pairing. Despite variable rates of UBP synthesis, the rates of extension remained problematic. While 6-aza substitution (d**NICS**) results in 2-fold reduced self-pair synthesis, it also results in a 2-fold increased rate of extension. Remarkably, replacing the oxygen of d**ICS** with sulfur (d**SICS**) results in an 80-fold increase in the rate of self-pair synthesis and a 4-fold increase in the rate of extension. The combination of both modifications in d**SNICS** retained the increased rate of synthesis but further increased the rate of extension 12-fold (k_{cat}/K_M = 2.2 × 10^4 M^{-1} min^{-1}). These results provided early hints as to the important but complicated role of the ortho substituent.

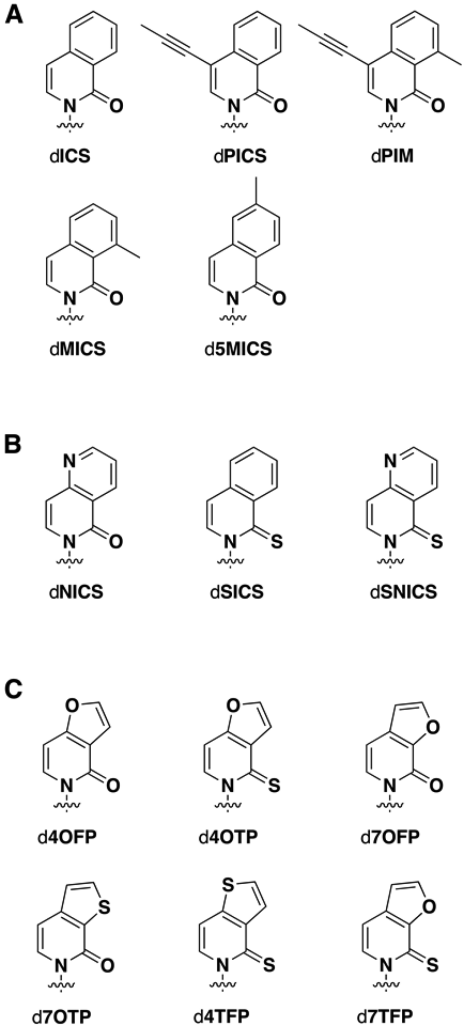


Figure 3. (A) Isocarbostyril analogs. (B) Heteroatom derivatized isocarbostyril analogs. (C) Furan and thiophene fused pyridone and thiopyridone analogs. Sugar and phosphate groups omitted for clarity.

A series of pyridones and thiopyridones fused to furan and thiophene rings in both meta- and para-linked orientations were investigated (Figure 3C).[15] UBP synthesis with these analogs was generally inefficient, with a para-linked furan appearing to be particularly detrimental. The thiopyridone triphosphate analogs are generally inserted more efficiently, but the effect of sulfur was less pronounced than with the isocarbostyril scaffold, and none of these analogs emerged as particularly promising.

Nucleoside analogs bearing azaindole scaffolds (Figure 4) were found to efficiently pair with various unnatural nucleotides in the template, and with reasonable selectively against the natural triphosphates when in the template, with the exception of d**P7AI**, which efficiently templates the insertion of dCTP.[16] However, while UBP synthesis again proved amenable to optimization, the resulting UBPs were generally refractory to extension, which by now had emerged as the greatest challenge to optimization.

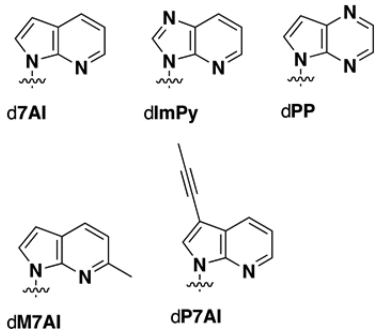


Figure 4. Azaindole analogs. Sugar and phosphate groups omitted for clarity.

A STRUCTURAL INTERLUDE

The dominant SAR that emerged from these first generation UBP candidates was that while the hydrophobicity and aromatic surface area of aromatic nucleobase derivatives appeared promising for the optimization of UBP synthesis, the resulting UBPs generally proved refractory to continued primer extension. The NMR structure of a duplex containing the d**PICS** self-pair (Figure 5A) was solved to better understand this SAR.[17] These studies revealed a generally unperturbed duplex structure, with the propargyl moieties of d**PICS** disposed as expected in the major groove; however,

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

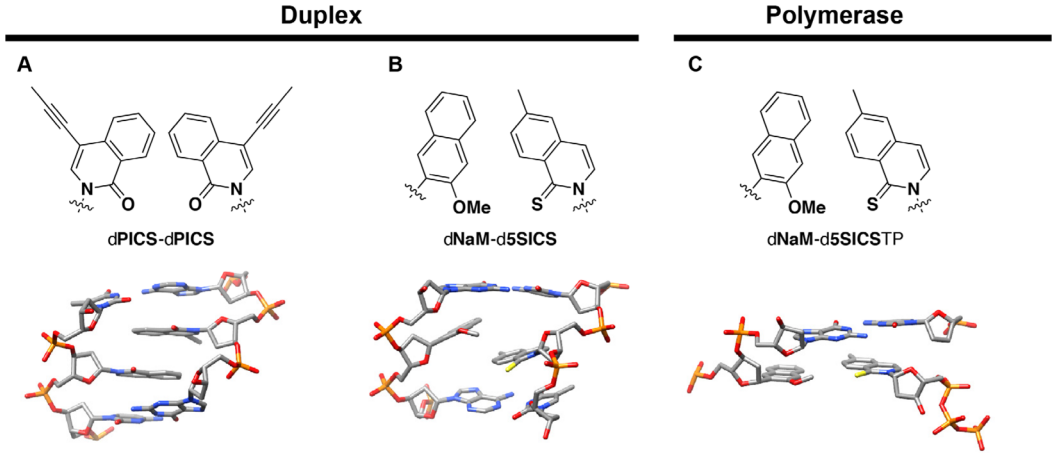


Figure 5. (A) Duplex structure of DNA containing the d**PICS**–d**PICS** UBP.[16] (B) Duplex structure of DNA containing the d**NaM**–d**5SICS** UBP.[36] (C) Structure of d**5SICS**TP paired opposite d**NaM** in the polymerase active site.[37] In chemical structures, sugar and phosphate groups omitted for clarity.

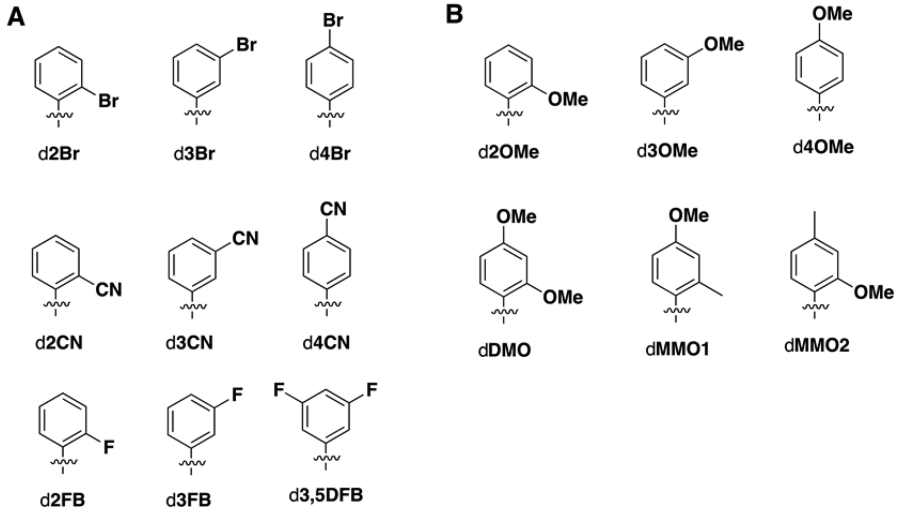


Figure 6. (A) Bromo-, cyano-, and fluoro-substituted benzene analogs. (B) Methoxy-substituted benzene analogs. Sugar and phosphate groups omitted for clarity.

considerable distortion was observed at the site of the UBP itself. Rather than adopting the canonical edge-on Watson–Crick geometry, the d**PICS** nucleobases interact via cross-strand intercalation. This intercalation appears to be driven by favorable stacking of the large aromatic surfaces of each d**PICS** nucleobase and has been observed with other nucleotides bearing nucleobase analogs with extended aromatic surface area.[18] We hypothesized that the same mode of pairing occurs at the primer terminus, which we envisioned would account for the efficient rates of UBP synthesis, but also the inefficient rates of continued extension, as deintercalation would be required to appropriately position the primer terminus.

CONTINUED PARTS OPTIMIZATION: SECOND GENERATION UBP CANDIDATES

Based on the SAR generated with the first generation analogs, we began testing whether smaller nucleobase analogs could be optimized for UBP synthesis, with the expectation that they would be less prone to cross-strand intercalate. These second generation efforts started with a more complete analysis of the simple benzene scaffold explored previously (Figure 2B). The parent analog, d**BEN**TP, is poorly recognized by Kf.[19] We then explored d**MM1**, d**MM2**, and d**MM3**, but found little improvement.[19] The efficiency of triphosphate insertion was progressively increased with d**DM**TP, d**DM2**TP, d**DM3**TP, d**DM4**TP, and d**DM5**TP, and insertion opposite dA was eliminated with d**TM**TP. Despite this progress with insertion, extension of primers terminating with these analogs remained inefficient.

We next explored heteroatom derivatization of these small scaffolds with bromo-, fluoro-, and cyano-substituents (Figure 6A). Bromo- and cyano-substituents tended to decrease the rate of mispairing with natural nucleotides,[20] but extension rates remained poor. A systematic analysis of fluoro substitution identified a single meta substituent (d**3FB**) as particularly interesting, with the resulting self-pair both synthesized and extended with at least moderate efficiency,[21] but further optimization efforts proved unproductive.

The situation improved with a family of methoxy-derivatized analogs (Figure 6B), and those possessing an ortho-methoxy substituent were the first to provide UBPs that were consistently extended with at least reasonable efficiency.[22] In fact, with d**MMO2** paired against d**TM** in the template, the resulting primer was extended with an efficiency that is only 36-

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394–403

fold lower than that of a natural base pair in the same sequence context. Mutation of the polymerase indicated that the increased extension resulted from the ability of the ortho-methoxy group to accept an H-bond, and similar substituents within the natural nucleobases are known to play a similar role.[23−25] Consistent with the need for an ortho H-bond acceptor at the primer terminus, d**TM** paired opposite d**MMO2** in the template is only extended poorly (k_{cat}/K_M = 6.3 × 10^3 M^{-1} min^{-1}). Although the d**3FB** self-pair was an exception, the SAR strongly suggested that an ortho H-bond acceptor was essential, and its inclusion emerged as a central design theme.

A variety of heterocyclic *N*- and *C*-nucleotides, which can also position an H-bond acceptor at the position ortho to the glycosidic linkage, were next examined[26−28] (Figures 7 and 8).

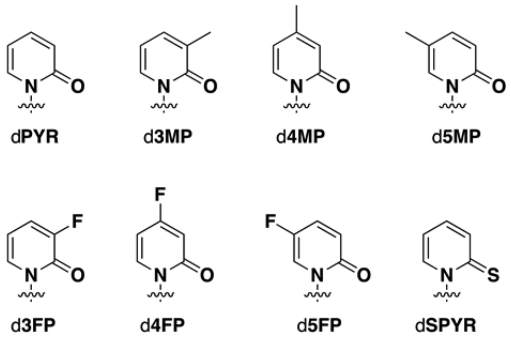


Figure 7. Derivatized monocyclic pyridone analogs. Sugar and phosphate groups omitted for clarity.

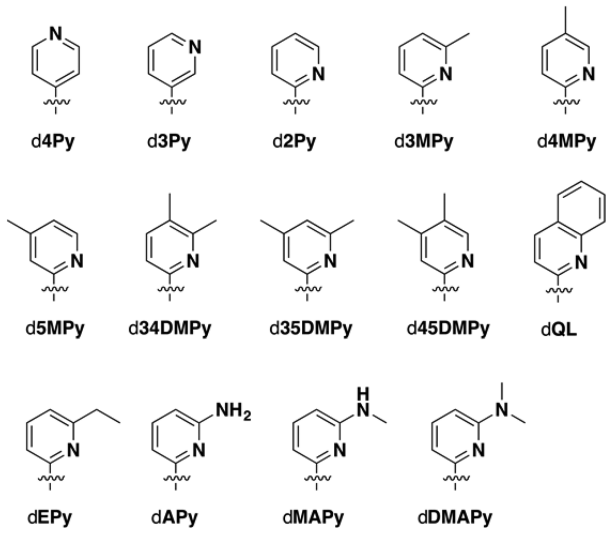


Figure 8. Pyridine and substituted pyridine analogs. Sugar and phosphate groups omitted for clarity.

The pyridone analog triphosphates were inserted with only marginal efficiency, but the resulting UBPs were extended with greater efficiency, consistent with the proposed role of the ortho H-bond acceptor. Conversion of d**PYR** to the corresponding thiopyridone (d**SPYR**) resulted in UBPs that were still reasonably well extended but synthesized less efficiently.[29] No pyridine analogs were efficiently inserted as triphosphates, but d**2Py** was better extended when at the primer terminus than was d**3Py** or d**4Py**, again consistent with the importance of an H-bond acceptor in the developing minor groove.[28] Problematically, however, the pyridine analogs paired more efficiently with dATP than any unnatural triphosphates when in the template, and no improvements were found with various alkyl or heteroatom substituents or with an increased aromatic surface (d**QL**).[30]

The most pronounced SAR generated from these second generation candidates was that efficient UBP synthesis requires a hydrophobic ortho substituent, while efficient extension requires the same substituent to be more hydrophilic, at least hydrophilic enough to act as an H-bond acceptor. This apparent physicochemical dichotomy challenged our confidence that both UBP synthesis and extension could be simultaneously optimized.

PARTS OPTIMIZATION: THIRD GENERATION UBP CANDIDATES

With no clear strategy to satisfy the conflicting demands of UBP synthesis and extension, we pivoted to a screen-based strategy. Two complementary screens were performed, one gel-based screen that focused on UBP incorporation and extension under steady state conditions, and one fluorescence-based plate screen that focused on the efficiency and fidelity of full length synthesis.[29] Remarkably, from 3600 candidate UBPs, both screens identified d**MMO2**−d**SICS** as the most promising. This UBP appeared to satisfy the physicochemical contradiction, because the sulfur of the d**SICS** thioamide is relatively hydrophobic but still able to accept an H-bond, while simple bond rotation allows the methoxy group of d**MMO2** to direct either a hydrophobic methyl group or the oxygen lone pairs into the developing minor groove.

The identification of d**MMO2**−d**SICS** reinvigorated our design efforts. Steady-state kinetics revealed that while d**MMO2** and d**SICS** were both incorporated and extended relatively efficiently (k_{cat}/K_M = 3.4 × 10^5 M^{-1} min^{-1} and 1.7 × 10^6 M^{-1} min^{-1}, respectively, with d**SICS** in the template, and 1.4 × 10^6 M^{-1} min^{-1} and 1.1 × 10^6 M^{-1} min^{-1}, respectively, with d**MMO2** in the template), replication fidelity was limited by d**SICS** self-pairing. Based on previous SAR, we explored the addition of a methyl group to the distal aromatic ring of d**SICS** to introduce steric interactions to disfavor self-pairing. These efforts identified d**5SICS**, and thus d**MMO2**−d**5SICS** emerged as our lead UBP.

We next turned to the optimization of d**MMO2**TP insertion opposite d**5SICS**, which was now the rate-limiting step of replication (Figure 9). Based on previous results that a meta-fluoro substituent or an expansion of aromatic surface area increased the efficiency of triphosphate incorporation, we explored the analogs d**5FM** and d**NaM** (Figure 9B).[31] Gratifyingly, opposite d**5SICS**, both d**5FM**TP and d**NaM**TP are more efficiently inserted than d**MMO2**TP. Although the former is limited by the synthesis and extension of the d**A**−d**5FM** mispair, the efficiency and fidelity of replicating DNA containing the d**NaM**−d**5SICS** UBP is excellent (k_{cat}/K_M = 5.0 × 10^6 M^{-1} min^{-1} and 3.7 × 10^7 M^{-1} min^{-1}, for insertion of d**NaM**TP opposite d**5SICS** and d**5SICS**TP opposite d**NaM**, respectively, and 1.2 × 10^6 M^{-1} min^{-1} and 2.7 × 10^6 M^{-1} min^{-1}, respectively, for each subsequent extension). In fact, d**NaM**−d**5SICS** was the first of our UBP candidates to be amplified in a standard PCR reaction with high efficiency and fidelity.[32] To determine whether d**NaM** is optimal for pairing with d**5SICS**, we further explored derivatization of the d**MMO2** scaffold with Cl, Br, I, Me, and Pr meta-substituents, but none improved replication.[33] We next examined 28 analogs with different para-substituents, and the SAR was augmented

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

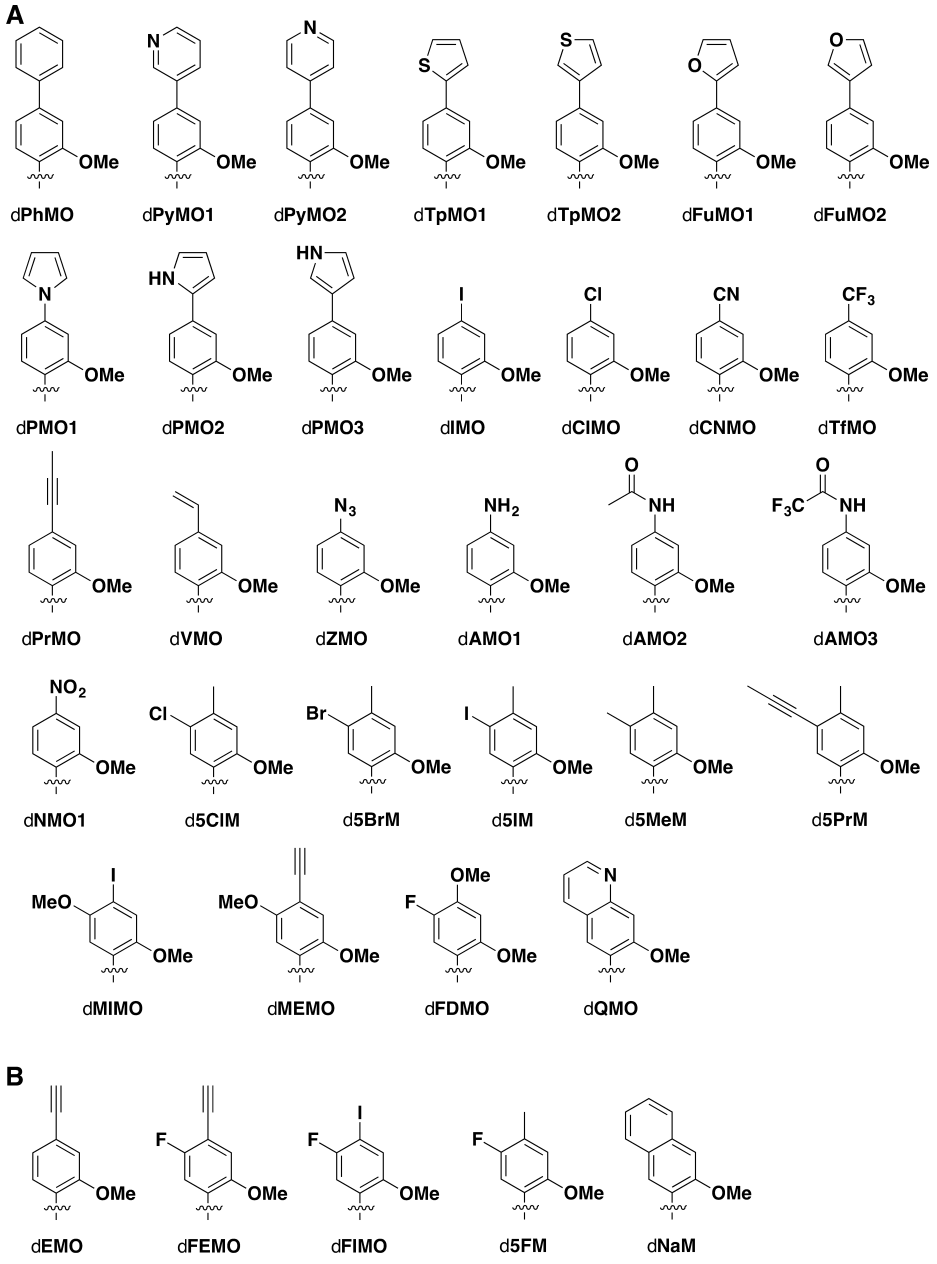


Figure 9. (A) Para- and meta-substituted d**MMO2** analogs. (B) Optimized d**MMO2** analogs. Sugar and phosphate groups omitted for clarity.

with PCR amplification assays using Taq, OneTaq, and KOD polymerases (Figure 9).[34,35] Several of the most promising analogs were further derivatized with a meta-fluoro or methoxy substituent. These efforts identified d**EMO**, d**FIMO**, and d**FEMO** as better partners for d**5SICS** than d**MMO2** (Figure 9B); however, none was more optimal in these *in vitro* assays than d**NaM**.

■ PARTS STANDARDIZATION

The successful development of a synthetic biology "part" includes not only its optimization for function, but also its standardization for function in different contexts, which here corresponds to recognition of the unnatural triphosphates by different DNA polymerases and within different local sequence contexts. To explore the extent of standardization of d**NaM**TP and d**5SICS**TP, we first examined the fidelity with which DNA containing the corresponding UBP was PCR-amplified in a variety of different sequence contexts using DeepVent, *Taq*, or Phusion polymerase, which demonstrated that retention of the UBP was near or in excess of 99% per doubling.[36] DNA containing a UBP within a 40-nt stretch of randomized natural nucleotides was PCR amplified 10^{24}-fold with *OneTaq*, and the products were analyzed by deep sequencing. A slight enrichment of the sequence 5′-GC**NaM** was observed, but this bias is no greater than that observed with natural sequences, demonstrating a sufficient level of standardization of the d**NaM**–d**5SICS** UBP for *in vivo* use.

■ A SECOND STRUCTURAL INTERLUDE

Having identified a family of promising UBPs, we again returned to a structural characterization. Collaborating with the Dwyer group, we first solved the structure of a DNA duplex containing d**MMO2**–d**5SICS**. Surprisingly, the UBP still formed a cross-strand intercalated structure,[37] and solution-

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394–403

state NOEs revealed that the d**NaM**−d**5SICS** UBP did as well (Figure 5B),[38] although significantly less so than the d**PICS** self-pair. While our interests centered around function and not structure, this mode of pairing with d**NaM**−d**5SICS** raised a perplexing question: because the UBP resembles a natural mispair more than a correct pair, how is it efficiently recognized by DNA polymerases, which are known to have evolved to reject triphosphates that form mispairs?

To address this question, we collaborated with the Marx group to solve the structures of the binary complex of KlenTaq DNA polymerase bound to a primer−template with d**NaM** in the templating position and with a primer terminating with a ddC, as well as the corresponding ternary complex with d**5SICS**TP bound. The data revealed that formation of the UBP drives the same large conformational change of the polymerase caused by the formation of a natural base pair (Figure 10)[39,40] and, remarkably, that the conformational

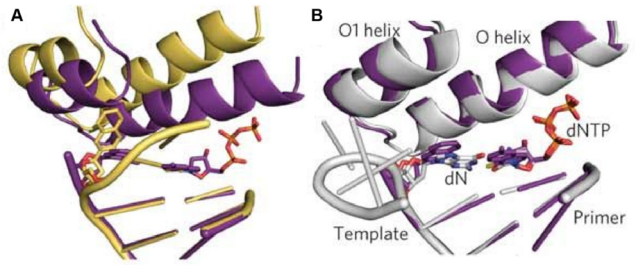


Figure 10. (A) Superimposition of the binary complex of KlenTaq polymerase bound to DNA with d**NaM** in the templating position in the open conformation (yellow) and the corresponding ternary complex bound to d**5SICS**TP in the closed conformation (purple). (B) Superimposition of ternary complex between KlenTaq polymerase, d**NaM** template DNA, and d**5SICS**TP (purple), or a natural dG template and dCTP (gray). Reproduced from ref 38. Copyright 2012 Nature Publishing Group.

change of the polymerase reciprocally drives a conformational change within the UBP, causing it to adopt an edge-to-edge paired natural Watson−Crick-like structure (Figure 5C). Thus, while a natural base pair is replicated with an induced-fit mechanism, the UBP is replicated with a similar, but subtly different, mutually induced-fit mechanism, providing a mechanistic explanation for its efficient replication. Nonetheless, after synthesis, the nascent UBP again adopts a cross-strand intercalated structure,[41] explaining the SAR data that identified a requirement for deintercalation prior to extension. At this point, we surmised that any further optimization of the UBP would require the careful optimization of intrastrand packing with neighboring natural nucleotides over cross-strand intercalation.

GROWING THE FAMILY OF *IN VITRO* REPLICATED UBPs

Based on the proposed mechanism of replication, we speculated that d**NaM**−d**5SICS** might be optimized by distal ring contraction and heteroatom derivatization of d**5SICS**, potentially favoring deintercalation, and thus extension, while preserving synthesis efficiency. Thus, we explored four derivatives with the distal ring replaced with para- or meta-linked thienyl, methyl furanyl, or methyl thienyl rings, as well as an additional derivative to explore fluorination at the meta position (Figure 11).[42] Gratifyingly, we found that d**TPT2**,

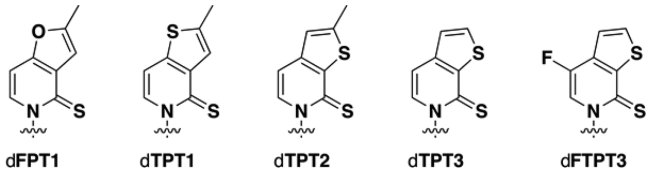


Figure 11. Distal ring-contracted d**5SICS** analogs. Sugar and phosphate groups omitted for clarity.

d**TPT3**, and d**FTPT3** form UBPs with d**NaM** that are more efficiently replicated within DNA than the d**NaM**-d**5SICS** UBP (as demonstrated by a pre-steady-state kinetic assay, as the steady-state rates were now limited by product dissociation[43]). The most efficiently replicated was d**NaM**−d**TPT3**, which thus emerged as our lead UBP.

At this point, we had explored several analogs of both d**MMO2** and d**5SICS** since the identification of d**MMO2**−d**SICS**, but we had never explored these analogs as partners with previous generation analogs. Thus, using a PCR-based screen, we examined the amplification of DNA containing 111 different unnatural nucleotides (resulting in approximately 6000 candidate UBPs) drawn from our now expanded set of analogs. While we found that d**NaM**−d**TPT3** is generally the most efficiently replicated of the UBPs examined, we identified seven additional and physicochemically distinct UBPs that are replicated significantly better than d**NaM**−d**5SICS** (Figure 12). Again drawing on established tenets of medicinal chemistry, these results are important, because our long-term goal of using the UBPs in a living SSO brings with it additional constraints, which may be differently satisfied by UBPs with differing physicochemical properties.

IN VIVO PERFORMANCE AND OPTIMIZATION

In 2014, we demonstrated that when d**NaM**TP and d**5SICS**TP are imported into *Escherichia coli* via transgenic expression of an algal nucleoside triphosphate transporter, they are used by cellular polymerases to replicate DNA containing the UBP.[5] However, unlike *in vitro* replication, replication in this SSO showed significant sequence biases, some of which were still observed with the d**NaM**−d**TPT3** UBP.[6] This is perhaps not surprising considering that the *in vivo* replication environment is distinct from the *in vitro* environment, which had been used to generate the SAR that drove UBP optimization and standardization. Thus, we screened 135 candidate UBPs, drawn from 91 unnatural triphosphates selected to cover the range of analogs that had been explored *in vitro*, for those that when added to the media were able to support high level retention of the corresponding UBP on a plasmid within the SSO.[44] Much of the SAR generated was consistent with that generated *in vitro*, but there were several interesting differences. In particular, the *in vivo* environment was somewhat more permissive to the nature of the ortho substituent that had proven so critical for *in vitro* replication (although the best UBPs still retained these H-bond acceptors). Perhaps more importantly, the *in vivo* screen identified four additional UBPs, each formed by pairing a d**NaM** analog opposite d**TPT3**, that are more efficiently replicated and with less sequence bias than d**NaM**−d**TPT3** (Figure 13), the most promising of which was d**CNMO**−d**TPT3**. While this UBP is at present the most promising lead for further SSO development, it is again the physicochemical diversity offered by a family of well replicated UBPs that is likely to prove most valuable as our efforts shift

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

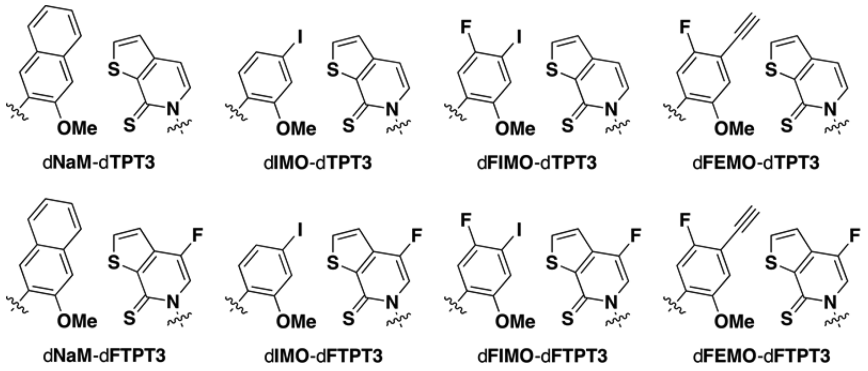


Figure 12. A family of well replicated d**NaM**−d**TPT3**-like UBPs. Sugar and phosphate groups omitted for clarity.

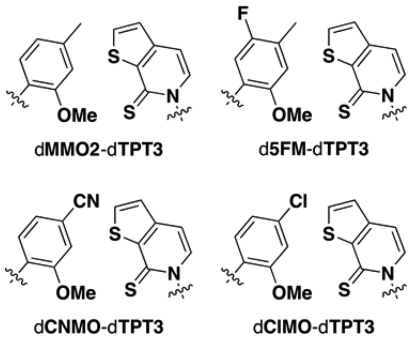


Figure 13. A family of UBPs optimized for *in vivo* expansion of the genetic alphabet. Sugar and phosphate groups omitted for clarity.

toward achieving *in vivo* transcription and translation, which we have only just begun to explore with d**NaM**−d**TPT3**.

CONCLUSIONS AND THE CHEMIST'S APPROACH TO SYNTHETIC BIOLOGY

We have used synthetic chemistry, coupled with the methods of medicinal chemistry, to develop a family of UBPs that function not only *in vitro* but also *in vivo* and have used them to create SSOs that can store more information in their DNA. The SARs elucidated from the examination of over 150 unnatural nucleotides have guided development and identified key elements that the unnatural nucleotides must possess, most clearly, an ortho substituent that is capable of providing a hydrophobic surface as well as an H-bond acceptor and a nucleobase surface that favors intrastrand packing over cross-strand intercalation. While this Account has recounted our efforts to optimize replication, we have also recently demonstrated that DNA containing the UBPs may be transcribed into RNA in an SSO and used during translation at the ribosome to produce proteins with noncanonical amino acids.[7] This lays the foundation for the creation of SSOs with forms and functions not available to their natural counterparts and thereby achieves a long-standing goal of synthetic biology.

At its core, synthetic biology aims to create parts that function within living cells, imparting them with novel attributes. While the tenets of the field were originally implicitly founded on the use of chemistry to create those parts, its modern incarnation has focused on the use of parts assembled from natural components or components intended to mimic their natural counterparts. This would seem justified by the eons of evolution that optimized the natural components for functioning in a cell, at least for a similar function. However, most natural components are recognized by or interact with multiple other components in every cell, possibly in unknown ways, and thus their introduction may have unintended consequences. While truly synthetic parts made by chemists do not benefit from eons of evolution, they are foreign to cells, possibly even drawing upon forces not used by their natural counterparts, and thus they may be more orthogonal and possibly easier to introduce and optimize without perturbation. In this case, optimization must proceed with less information, because less is known about how the parts might interact with their biological targets, and the methodology and lessons of medicinal chemistry, which conceptually face essentially the same challenges, provide the blueprint for success. If the combination of synthetic and medicinal-like chemistry can produce molecules that function alongside those evolved by nature for the most central of its processes, to store and retrieve information, then this approach is likely to be capable of discovering molecules that effectively participate in any biological process, potentially opening a new vista for chemists in synthetic biology.

AUTHOR INFORMATION

Corresponding Author

*E-mail: floyd@scripps.edu.

ORCID

Aaron W. Feldman: 0000-0002-9495-2357
Floyd E. Romesberg: 0000-0001-6317-1315

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

Financial support was provided by the National Institutes of Health (GM060005 and GM118178 to F.E.R.) and the National Science Foundation Graduate Research Fellowship (DGE-1346837 to A.W.F.)

Notes

The authors declare the following competing financial interest(s): F.E.R. has a financial interest (shares) in Synthorx Inc., a company that has commercial interests in the UBP. The other author declares no other competing financial interests.

Biographies

Aaron W. Feldman received his A.B. in Chemistry from Connecticut College in 2013. He is currently a graduate student in the Romesberg

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

lab working to chemically optimize semi-synthetic organisms with expanded genetic alphabets.

Floyd E. Romesberg received his Ph.D. in Chemistry from Cornell University in 1995. He is currently Professor of Chemistry at The Scripps Research Institute.

■ ACKNOWLEDGMENTS

We thank former and current members of the Romesberg group, as well as those of our collaborators in the Dwyer and Marx groups for their contributions to the research highlighted in this Account.

■ REFERENCES

(1) Leduc, S. *The Mechanisms of Life*; Rebman Company: New York, 1911.
(2) Benner, S. A.; Karalkar, N. B.; Hoshika, S.; Laos, R.; Shaw, R. W.; Matsuura, M.; Fajardo, D.; Moussatche, P. Alternative Watson-Crick synthetic genetic systems. *Cold Spring Harbor Perspect. Biol.* **2016**, *8*, a023770.
(3) Moran, S.; Ren, R. X.; Kool, E. T. A thymidine triphosphate shape analog lacking Watson-Crick pairing ability is replicated with high sequence selectivity. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 10506−10511.
(4) Hirao, I.; Kimoto, M.; Yamashige, R. Natural versus artificial creation of base pairs in DNA: origin of nucleobases from the perspectives of unnatural base pair studies. *Acc. Chem. Res.* **2012**, *45*, 2055−2065.
(5) Malyshev, D. A.; Dhami, K.; Lavergne, T.; Chen, T.; Dai, N.; Foster, J. M.; Correa, I. R., Jr.; Romesberg, F. E. A semi-synthetic organism with an expanded genetic alphabet. *Nature* **2014**, *509*, 385−388.
(6) Zhang, Y.; Lamb, B. M.; Feldman, A. W.; Zhou, A. X.; Lavergne, T.; Li, L.; Romesberg, F. E. A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 1317−1322.
(7) Zhang, Y.; Ptacin, J. L.; Fischer, E. C.; Aerni, H. R.; Caffaro, C. E.; San Jose, K.; Feldman, A. W.; Turner, C. R.; Romesberg, F. E. A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* **2017**, DOI: 10.1038/nature24659.
(8) Hwang, G. T.; Hari, Y.; Romesberg, F. E. The effects of unnatural base pairs and mispairs on DNA duplex stability and solvation. *Nucleic Acids Res.* **2009**, *37*, 4757−4763.
(9) Petruska, J.; Goodman, M. F.; Boosalis, M. S.; Sowers, L. C.; Cheong, C.; Tinoco, I., Jr Comparison between DNA melting thermodynamics and DNA polymerase fidelity. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 6252−6256.
(10) Ogawa, A. K.; Wu, Y. Q.; McMinn, D. L.; Liu, J. Q.; Schultz, P. G.; Romesberg, F. E. Efforts toward the expansion of the genetic alphabet: Information storage and replication with unnatural hydrophobic base pairs. *J. Am. Chem. Soc.* **2000**, *122*, 3274−3287.
(11) Ogawa, A. K.; Wu, Y. Q.; Berger, M.; Schultz, P. G.; Romesberg, F. E. Rational design of an unnatural base pair with increased kinetic selectivity. *J. Am. Chem. Soc.* **2000**, *122*, 8803−8804.
(12) Berger, M.; Wu, Y.; Ogawa, A. K.; McMinn, D. L.; Schultz, P. G.; Romesberg, F. E. Universal bases for hybridization, replication and chain termination. *Nucleic Acids Res.* **2000**, *28*, 2911−2914.
(13) McMinn, D. L.; Ogawa, A. K.; Wu, Y.; Liu, J.; Schultz, P. G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: DNA polymerase recognition of a highly stable, self-pairing hydrophobic base. *J. Am. Chem. Soc.* **1999**, *121*, 11585−11586.
(14) Yu, C.; Henry, A. A.; Romesberg, F. E.; Schultz, P. G. Polymerase recognition of unnatural base pairs. *Angew. Chem., Int. Ed.* **2002**, *41*, 3841−3844.
(15) Henry, A. A.; Yu, C.; Romesberg, F. E. Determinants of unnatural nucleobase stability and polymerase recognition. *J. Am. Chem. Soc.* **2003**, *125*, 9638−9646.
(16) Wu, Y.; Ogawa, A. K.; Berger, M.; McMinn, D. L.; Schultz, P. G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: optimization of interbase hydrophobic interactions. *J. Am. Chem. Soc.* **2000**, *122*, 7621−7632.
(17) Matsuda, S.; Fillo, J. D.; Henry, A. A.; Rai, P.; Wilkens, S. J.; Dwyer, T. J.; Geierstanger, B. H.; Wemmer, D. E.; Schultz, P. G.; Spraggon, G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: Structure and replication of unnatural base pairs. *J. Am. Chem. Soc.* **2007**, *129*, 10466−10473.
(18) Malyshev, D. A.; Romesberg, F. E. The expanded genetic alphabet. *Angew. Chem., Int. Ed.* **2015**, *54*, 11930−11944.
(19) Matsuda, S.; Henry, A. A.; Romesberg, F. E. Optimization of unnatural base pair packing for polymerase recognition. *J. Am. Chem. Soc.* **2006**, *128*, 6369−6375.
(20) Hwang, G. T.; Romesberg, F. E. Substituent effects on the pairing and polymerase recognition of simple unnatural base pairs. *Nucleic Acids Res.* **2006**, *34*, 2037−2045.
(21) Henry, A. A.; Olsen, A. G.; Matsuda, S.; Yu, C. Z.; Geierstanger, B. H.; Romesberg, F. E. Efforts to expand the genetic alphabet: Identification of a replicable unnatural DNA self-pair. *J. Am. Chem. Soc.* **2004**, *126*, 6923−6931.
(22) Matsuda, S.; Leconte, A. M.; Romesberg, F. E. Minor groove hydrogen bonds and the replication of unnatural base pairs. *J. Am. Chem. Soc.* **2007**, *129*, 5551−5557.
(23) Spratt, T. E. Identification of hydrogen bonds between *Escherichia coli* DNA polymerase I (Klenow fragment) and the minor groove of DNA by amino acid substitution of the polymerase and atomic substitution of the DNA. *Biochemistry* **2001**, *40*, 2647−2652.
(24) Morales, J. C.; Kool, E. T. Minor groove interactions between polymerase and DNA: more essential to replication than Watson−Crick hydrogen bonds? *J. Am. Chem. Soc.* **1999**, *121*, 2323−2324.
(25) Meyer, A. S.; Blandino, M.; Spratt, T. E. *Escherichia coli* DNA polymerase I (Klenow fragment) uses a hydrogen-bonding fork from Arg668 to the primer terminus and incoming deoxynucleotide triphosphate to catalyze DNA replication. *J. Biol. Chem.* **2004**, *279*, 33043−33046.
(26) Leconte, A. M.; Matsuda, S.; Hwang, G. T.; Romesberg, F. E. Efforts towards expansion of the genetic alphabet: pyridone and methyl pyridone nucleobases. *Angew. Chem., Int. Ed.* **2006**, *45*, 4326−4329.
(27) Hwang, G. T.; Leconte, A. M.; Romesberg, F. E. Polymerase recognition and stability of fluoro-substituted pyridone nucleobase analogues. *ChemBioChem* **2007**, *8*, 1606−1611.
(28) Kim, Y.; Leconte, A. M.; Hari, Y.; Romesberg, F. E. Stability and polymerase recognition of pyridine nucleobase analogues: Role of minor-groove H-bond acceptors. *Angew. Chem., Int. Ed.* **2006**, *45*, 7809−7812.
(29) Leconte, A. M.; Hwang, G. T.; Matsuda, S.; Capek, P.; Hari, Y.; Romesberg, F. E. Discovery, characterization, and optimization of an unnatural base pair for expansion of the genetic alphabet. *J. Am. Chem. Soc.* **2008**, *130*, 2336−2343.
(30) Hari, Y.; Hwang, G. T.; Leconte, A. M.; Joubert, N.; Hocek, M.; Romesberg, F. E. Optimization of the pyridyl nucleobase scaffold for polymerase recognition and unnatural base pair replication. *ChemBioChem* **2008**, *9*, 2796−2799.
(31) Seo, Y. J.; Hwang, G. T.; Ordoukhanian, P.; Romesberg, F. E. Optimization of an unnatural base pair toward natural-like replication. *J. Am. Chem. Soc.* **2009**, *131*, 3246−3252.
(32) Malyshev, D. A.; Dhami, K.; Quach, H. T.; Lavergne, T.; Ordoukhanian, P.; Torkamani, A.; Romesberg, F. E. Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 12005−12010.
(33) Seo, Y. J.; Romesberg, F. E. Major groove derivatization of an unnatural base pair. *ChemBioChem* **2009**, *10*, 2394−2400.
(34) Lavergne, T.; Malyshev, D. A.; Romesberg, F. E. Major groove substituents and polymerase recognition of a class of predominantly

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403

hydrophobic unnatural base pairs. *Chem. - Eur. J.* **2012**, *18*, 1231−1239.
(35) Lavergne, T.; Degardin, M.; Malyshev, D. A.; Quach, H. T.; Dhami, K.; Ordoukhanian, P.; Romesberg, F. E. Expanding the scope of replicable unnatural DNA: stepwise optimization of a predominantly hydrophobic base pair. *J. Am. Chem. Soc.* **2013**, *135*, 5408−5419.
(36) Malyshev, D. A.; Seo, Y. J.; Ordoukhanian, P.; Romesberg, F. E. PCR with an expanded genetic alphabet. *J. Am. Chem. Soc.* **2009**, *131*, 14620−14621.
(37) Malyshev, D. A.; Pfaff, D. A.; Ippoliti, S. I.; Hwang, G. T.; Dwyer, T. J.; Romesberg, F. E. Solution structure, mechanism of replication, and optimization of an unnatural base pair. *Chem. - Eur. J.* **2010**, *16*, 12650−12659.
(38) Betz, K.; Malyshev, D. A.; Lavergne, T.; Welte, W.; Diederichs, K.; Dwyer, T. J.; Ordoukhanian, P.; Romesberg, F. E.; Marx, A. KlenTaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. *Nat. Chem. Biol.* **2012**, *8*, 612−614.
(39) Rothwell, P. J.; Waksman, G. Structure and mechanism of DNA polymerases. *Adv. Protein Chem.* **2005**, *71*, 401−440.
(40) Wu, E. Y.; Beese, L. S. The structure of a high fidelity DNA polymerase bound to a mismatched nucleotide reveals an ″ajar″ intermediate conformation in the nucleotide selection mechanism. *J. Biol. Chem.* **2011**, *286*, 19758−19767.
(41) Betz, K.; Malyshev, D. A.; Lavergne, T.; Welte, W.; Diederichs, K.; Romesberg, F. E.; Marx, A. Structural insights into DNA replication without hydrogen bonds. *J. Am. Chem. Soc.* **2013**, *135*, 18637−18643.
(42) Li, L.; Degardin, M.; Lavergne, T.; Malyshev, D. A.; Dhami, K.; Ordoukhanian, P.; Romesberg, F. E. Natural-like replication of an unnatural base pair for the expansion of the genetic alphabet and biotechnology applications. *J. Am. Chem. Soc.* **2014**, *136*, 826−829.
(43) Morris, S. E.; Feldman, A. W.; Romesberg, F. E. Synthetic biology parts for the storage of increased genetic information in cells. *ACS Synth. Biol.* **2017**, *6*, 1834−1840.
(44) Feldman, A. W.; Dien, V. T.; Romesberg, F. E. Chemical stabilization of unnatural nucleotide triphosphates for the in vivo expansion of the genetic alphabet. *J. Am. Chem. Soc.* **2017**, *139*, 2464−2467.

DOI: 10.1021/acs.accounts.7b00403
Acc. Chem. Res. 2018, 51, 394−403