Nucleic Acids Research, 2014, Vol. 42, No. 16 ***10235–10244***
doi: 10.1093/nar/gku715

Systematic exploration of a class of hydrophobic unnatural base pairs yields multiple new candidates for the expansion of the genetic alphabet

Kirandeep Dhami[1], Denis A. Malyshev[1], Phillip Ordoukhanian[2], Tomáš Kubelka[3], Michal Hocek[3,4] and Floyd E. Romesberg[1,*]

[1]Department of Chemistry, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA, [2]Center for Protein and Nucleic Acid Research, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA, [3]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo nam. 2, CZ-16610 Prague 6, Czech Republic and [4]Department of Organic Chemistry, Faculty of Science, Charles University in Prague, Hlavova 8, CZ-12843 Prague 2, Czech Republic

Received June 4, 2014; Revised July 22, 2014; Accepted July 23, 2014

ABSTRACT

We have developed a family of unnatural base pairs (UBPs), which rely on hydrophobic and packing interactions for pairing and which are well replicated and transcribed. While the pair formed between d5SICS and dNaM (d5SICS-dNaM) has received the most attention, and has been used to expand the genetic alphabet of a living organism, recent efforts have identified dTPT3-dNaM, which is replicated with even higher fidelity. These efforts also resulted in more UBPs than could be independently analyzed, and thus we now report a PCR-based screen to identify the most promising. While we found that dTPT3-dNaM is generally the most promising UBP, we identified several others that are replicated nearly as well and significantly better than d5SICS-dNaM, and are thus viable candidates for the expansion of the genetic alphabet of a living organism. Moreover, the results suggest that continued optimization should be possible, and that the putatively essential hydrogen-bond acceptor at the position *ortho* to the glycosidic linkage may not be required. These results clearly demonstrate the generality of hydrophobic forces for the control of base pairing within DNA, provide a wealth of new structure–activity relationship data and importantly identify multiple new candidates for *in vivo* evaluation and further optimization.

INTRODUCTION

Expansion of the genetic alphabet by development of a replicable unnatural base pair (UBP) has attracted significant attention (1–7) since report of the first efforts in 1989 (8). For over a decade, we have explored the use of hydrophobic and packing forces to drive the stable and selective pairing of unnatural nucleotides in DNA and during replication and transcription. Our initial work focused on the replacement of the natural purine or pyrimidine nucleobases with predominantly hydrophobic analogs based on benzene-, naphthalene-, isocarbostiryl-, pyridine- and pyridone-scaffolds (9). However, the number of candidate UBPs formed by these nucleotides soon exceeded the number that could be analyzed individually, and as a result, we conducted a screen wherein 3600 candidate UBPs were analyzed (10). This screen identified the pair formed between d**SICS** and d**MMO2** (d**SICS**-d**MMO2**), which upon optimization yielded d**5SICS**-d**MMO2** (Figure 1), whose relatively efficient replication by a variety of DNA polymerases (11) validated the use of hydrophobic and packing forces instead of the canonical Watson–Crick hydrogen-bonds (H-bonds) that underlie the replication of natural base pairs. In addition, one of clearest structure–activity relationships (SARs) to emerge from these studies was the apparent importance of an H-bond acceptor positioned *ortho* to the glycosidic linkage (10,12,13), which as with natural DNA (14–16), was thought to mediate the formation of a critical H-bond with a polymerase donor.

Our efforts to optimize the UBP then turned to improving d**MMO2** as a partner for d**5SICS**, eventually yielding d**5SICS**-d**NaM** (Figure 1). d**5SICS**-d**MMO2** and especially d**5SICS**-d**NaM** are replicated (2,10,17) and transcribed (18) sufficiently well for many applications, and we have used linker-modified versions to enzymatically synthesize site-specifically labeled DNA and RNA (4,19). Most importantly, we have incorporated d**5SICS**-d**NaM** into a plasmid that is stably propagated in *Escherichia coli*, creating the first semi-synthetic organism with an expanded genetic alphabet (20). Nonetheless, the demands of *in vivo* replication in

*To whom correspondence should be addressed. Tel: +1 858 784 7290; Fax: +1 858 784 7472; Email: floyd@scripps.edu

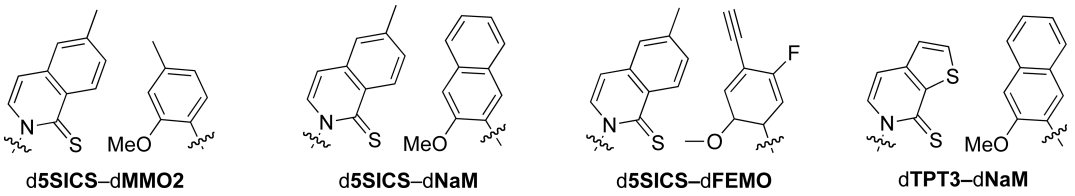

Figure 1. The most promising UBPs previously identified. Sugar and phosphate backbone are omitted for clarity.

different organisms and in all possible sequence contexts, including those containing multiple UBPs, is likely to require further optimization, which has proceeded, at least in part, based on the structure of the UBP formed in the active site of a DNA polymerase (3,21). Additional analogs of both d**5SICS** and d**NaM** were synthesized that in all cases maintained the putatively essential H-bond acceptor positioned *ortho* to the glycosidic bond. These efforts yielded d**5SICS**-d**FEMO** (22), and ultimately d**TPT3**-d**NaM** (23) (Figure 1). Although d**5SICS**-d**FEMO** is replicated with an efficiency and fidelity similar to that of d**5SICS**-d**NaM**, its propinyl group provides a natural site for post-amplification derivatization and thus for site-specific DNA labeling. In contrast, when incorporated into DNA d**TPT3**-d**NaM** is replicated better than either d**5SICS**-d**MMO2** or d**5SICS**-d**NaM**, with rates approaching those of a natural base pair.

Despite the efficient replication of DNA containing d**TPT3**-d**NaM**, it was uncertain whether it was the most promising UBP formed between the nucleotides that had been synthesized. This is because the optimization efforts again resulted in the synthesis of too many nucleotide analogs to analyze all possible combinations individually. Here, we report a screen of a library of 111 unnatural nucleotides (resulting in ∼6000 candidate UBPs) drawn from our complete set of analogs, including those synthesized after the identification d**5SICS**-d**MMO2**, which are structurally more homologous to either d**5SICS** or d**NaM**, as well as those synthesized before the identification of d**5SICS**-d**MMO2**, which are more structurally diverse. However, unlike our previously reported screen, which was based on the steady-state synthesis of a single strand of DNA, the current screen relies on polymerase chain reaction (PCR) amplification. This approach was selected because PCR is an important practical tool with which any identified UBPs should be immediately compatible, and further it has allowed us to screen for fidelity in a straightforward manner, by sequencing the amplification products. We found that d**TPT3**-d**NaM** is, in general, the most efficiently replicated of the UBPs examined; however, we also identified seven additional and structurally distinct UBPs that are better replicated than d**5SICS**-d**NaM** and should thus be sufficiently well replicated to underlie the expansion of an organism's genetic alphabet. In addition, we identified two UBPs that are reasonably well replicated despite one constituent nucleobase lacking the putatively essential *ortho* H-bond acceptor, which challenges the assumption, at least for UBPs, that the H-bond acceptor is required for efficient replication. Finally, additional SAR data were generated that should facilitate the further optimization of this class of UBP.

MATERIALS AND METHODS

General

The triphosphates of the α6 group were prepared from the previously reported nucleosides (24) according to literature procedure (25) (See Synthetic Methods and Spectra in Supplementary Data and Supplementary Table S1). The purity of all other triphosphates was confirmed by MALDI-TOF and UV-VIS. Taq and OneTaq DNA polymerases were purchased from New England Biolabs (Ipswich, MA, USA). A mixture of dNTPs was purchased from Fermentas (Glen Burnie, MD, USA). SYBR Green I Nucleic Acid Gel Stain (10 000×) was purchased from Life Technologies (Carlsbad, CA, USA). The synthesis of the DNA templates, D8 (2), used for screening rounds 1–5, and D6 (26), used for all other amplifications, was described previously; sequences of templates are provided in Supplementary Table S2. Sanger sequencing was carried out as described previously (2). Raw Sanger sequencing traces were used to determine the percent retention of the UBPs, which was converted to fidelity per doubling, as described previously (2,26) and in the Supplementary Data.

Screen PCR assay conditions

All PCR amplifications were performed in a CFX Connect Real-Time PCR Detection System (Bio-Rad), in a total volume of 25 μl using the following conditions: 1× OneTaq reaction buffer, 0.5× Sybr Green I, $MgSO_4$ adjusted to 4.0 mM, 0.2 mM of each dNTP, 50 μM of each unnatural triphosphate, 1 μM of Primer1 and Primer2 (See Supplementary Table S2) and 0.02 U/μl of the DNA polymerase. Other conditions specific for each round of screening are described in Supplementary Table S3. Amplified products were purified using DNA Clean and Concentrator-5 spin columns from Zymo Research (Irvine, CA, USA). After purification, the PCR products were sequenced on a 3730 DNA Analyzer (Applied Biosystems) to determine the retention of the UBP as described in the Supplementary Material. Fidelity was characterized from UBP retention as determined by sequencing with Primer1 on a 3730 DNA Analyzer (Applied Biosystems).

Specific PCR assay conditions

PCR with the most promising UBPs was carried out with the conditions as described in Supplementary Table S3. PCR products were further purified on 2% agarose gels, followed by single band excision and subsequent clean up using the Zymo Research Zymoclean Gel DNA Recovery Kit. After elution with 20 μl of water, the DNA concentration

was measured using fluorescent dye binding (Quant-iT dsDNA HS Assay kit, Life Technologies), and purified amplicons were sequenced in triplicate with both Primer1 and Primer2 to determine UBP retention and thus amplification fidelity (see Supplementary Figures S1–S3). Amplification of DNA containing the pairs involving analogs of group α6 was performed with OneTaq polymerase under the following thermal cycling conditions: initial denaturation at 96°C for 1 min; 16 cycles of 96°C for 10 s, 60°C for 15 s, 68°C for 1 min. Fidelity was determined by sequencing amplicons in the Primer1 direction in triplicate (Supplementary Figure S4). Amplification of DNA containing the UBPs formed between d**TPT3** and d**2MN** or d**DM2** was performed using OneTaq or Taq polymerases for 16 cycles under the following thermal cycling conditions: (i) OneTaq: initial denaturation at 96°C for 1 min, 96°C for 10 s, 60°C for 15 s, 68°C for 1 min; or (ii) Taq: initial denaturation at 96°C for 1 min, 96°C for 5 s, 60°C for 5 s, 68°C for 10 s. Fidelity was determined by sequencing amplicons in the Primer1 direction in triplicate (Supplementary Figure S5).

RESULTS

To screen for well replicated UBPs, unnatural deoxynucleoside triphosphates were grouped for analysis into either d**MMO2**/d**NaM**- or d**5SICS**/d**TPT3**-like analogs, although the distinction is not completely clear in all cases. In total, 80 d**MMO2**/d**NaM** analogs were grouped into 12 'α groups' (α1–α12; Figure 2), and 31 d**5SICS**/d**TPT3** analogs were grouped into six 'β groups' (β1–β6; Figure 3). Note that the group designations used here should not be confused with anomer designation (all analogs examined are β glycosides). In addition, to increase the SAR content of the screen, seven previously reported nucleoside analogs (d**TOK576**–d**TOK588**) with substituted pyridyl nucleobases (24) were phosphorylated as described in Supplementary Data, and included as group α6. For screening, a 134-mer single-stranded DNA template containing a centrally located d**NaM** (which has been used previously and is referred to as D8 (2)) was PCR amplified in the presence of the natural triphosphates (200 μM each), all pairwise combinations of an α and a β triphosphate group (50 μM each), and 0.02 U/μl DNA polymerase. During the first round of PCR, d**NaM** templates the incorporation of a β analog and is then replaced by an α analog when the original strand is copied in the second round, with the resulting UBP amplified in subsequent rounds. The amplification product of each reaction was analyzed by Sanger sequencing (Supplementary Data). As reported previously, the presence of an unnatural nucleotide results in the abrupt termination of the sequencing chromatogram, allowing the level of UBP retention to be quantified by the amount of read through (2,26). The percentage of UBP retained in the DNA after amplification during each round of screening is shown in Figure 4.

The first round of screening employed 0.1 ng of template and 16 cycles of amplification under relatively permissive conditions that included OneTaq polymerase and a 1 min extension time. For our purposes, OneTaq is considered permissive because it is a mixture of Taq (a family A polymerase (27,28)) and Deep Vent (a family B polymerase (27,28)), with the latter possessing exonucleotidic proofreading that allows for the excision of an incorrectly incorporated triphosphate. Under these conditions, only the pairs involving group β5 or β6 showed high retention.

The combinations of β5 or β6 and the α groups that showed the highest retention were progressed to a second round of screening, wherein they were divided into smaller groups (denoted by a, b or c; Figures 2 and 3). High retention (≥97%) was observed with β5a and α2c, α9a, α9c, α10a, α10c, α12b or α12c; with β5b and α9a, α9b, α10c or α12b and with β6b and α10c. Moderate retention (84–96%) was observed with β5a and α1a, α1b, α6a, α9b, α10b or α12a; β5b and α1a, α1b, α2c, α6a, α9c, α10a, α12a or α12c; β6a and α1b or α10c and β6b and α1a, α6a, α9α–c, α10a, α10b or α12a–c.

For a third round of screening, α analogs were analyzed in groups of only one to three compounds, and group β6a was subdivided into its two constituent triphosphates, d**TPT1**TP and d**FPT1**TP. The highest retention (≥90%) was observed with β5a and α1a, α2cII, α9a–c, α10aI, α10aII, α10c, α12b or d**TfMO**TP; β5b and α9a, α9c or α10c; d**FPT1**TP and α10aI and β6b and α1a, α9a–c, α10aI, α10aII, α10c, α12b, d**NMO**TP, d**TfMO**TP or d**CNMO**TP. Only slightly less retention (80–89%) was seen with β5a and α2cI, α12a, d**NMO**TP, d**QMO**TP or d**TOK587**TP; β5b and α1a, α2cII, α10aI, α10aII, α12b or d**TOK587**TP; d**FPT1**TP and α10c and β6b and α12a, d**QMO**TP, d**FuMO1**TP or d**TOK587**TP.

For a fourth round of screening, all of the α derivatives were analyzed as individual triphosphates, with the exception of α9b and α9c, which remained grouped. The highest retention (≥91%) was observed with β5a and α9b, α9c, d**FIMO**TP, d**IMO**TP, d**FEMO**TP, d**MMO2**TP, d**2OMe**TP, d**DMO**TP, d**5FM**TP, d**NaM**TP, d**VMO**TP, d**ZMO**TP, d**ClMO**TP, d**TfMO**TP, d**QMO**TP, d**2MN**TP, d**DM2**TP or d**TOK587**TP; β5b and α9b, α9c, d**FIMO**TP, d**IMO**TP, d**FEMO**TP, d**NaM**TP, d**ZMO**TP, d**ClMO**TP, d**QMO**TP, d**MM1**TP, d**DM2**TP or d**TOK587**TP; β6 analog d**FPT1**TP and α analogs d**2OMe**TP or d**NaM**TP and β6b and α9b, α9c, d**FIMO**TP, d**IMO**TP, d**FEMO**TP, d**MMO2**TP, d**DMO**TP, d**TMO**TP, d**NMO**TP, d**5FM**TP, d**NaM**TP, d**VMO**TP, d**ZMO**TP, d**ClMO**TP, d**TfMO**TP, d**QMO**TP, d**CNMO**TP, d**2MN**TP, d**TOK587**TP or d**FuMO2**TP.

To increase the stringency of the screen, a fifth round was performed with Taq polymerase instead of OneTaq, as it lacks exonuclease proofreading activity and thus increases the sensitivity to mispair synthesis. This round also separated all remaining α and β groups into individual triphosphates. The highest retention (≥90%) was seen with d**SICS**TP and d**NaM**TP; d**SNICS**TP and d**NaM**TP; d**TPT2**TP and d**FDMO**TP; d**TPT3**TP and d**FIMO**TP, d**IMO**TP or d**NaM**TP and d**FTPT3**TP and d**FIMO**TP, d**IMO**TP, d**FEMO**TP, d**NMO**TP, d**NaM**TP, d**ClMO**TP, d**TfMO**TP or d**CNMO**TP.

To better differentiate between the UBPs, we progressed the 62 most promising candidate UBPs to a sixth round of screening in which the template concentration was decreased 10-fold (to 10 pg) to allow for greater amplification, and thereby afford greater discrimination, and the template was changed to D6 (26), where the three flanking nucleotides on either side of the unnatural nucleotide are randomized among the natural

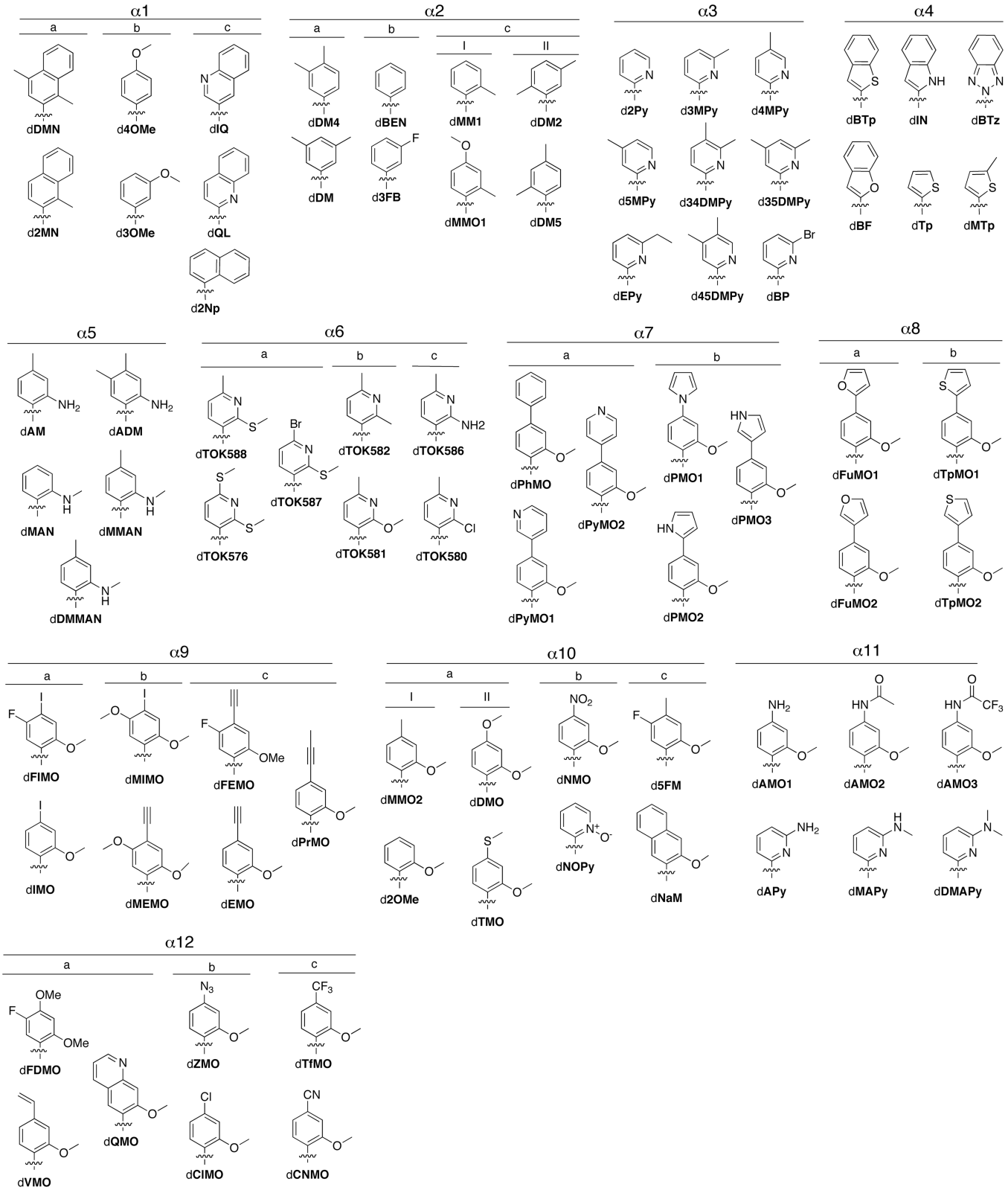

Figure 2. α group unnatural deoxynucleoside triphosphates. d**2OMe** and d**MMO1** were moved from Group α1 to the groups indicated after the first round of screening. Sugar and phosphate backbone are omitted for clarity. References for each compound are provided in Supplementary Table S6.

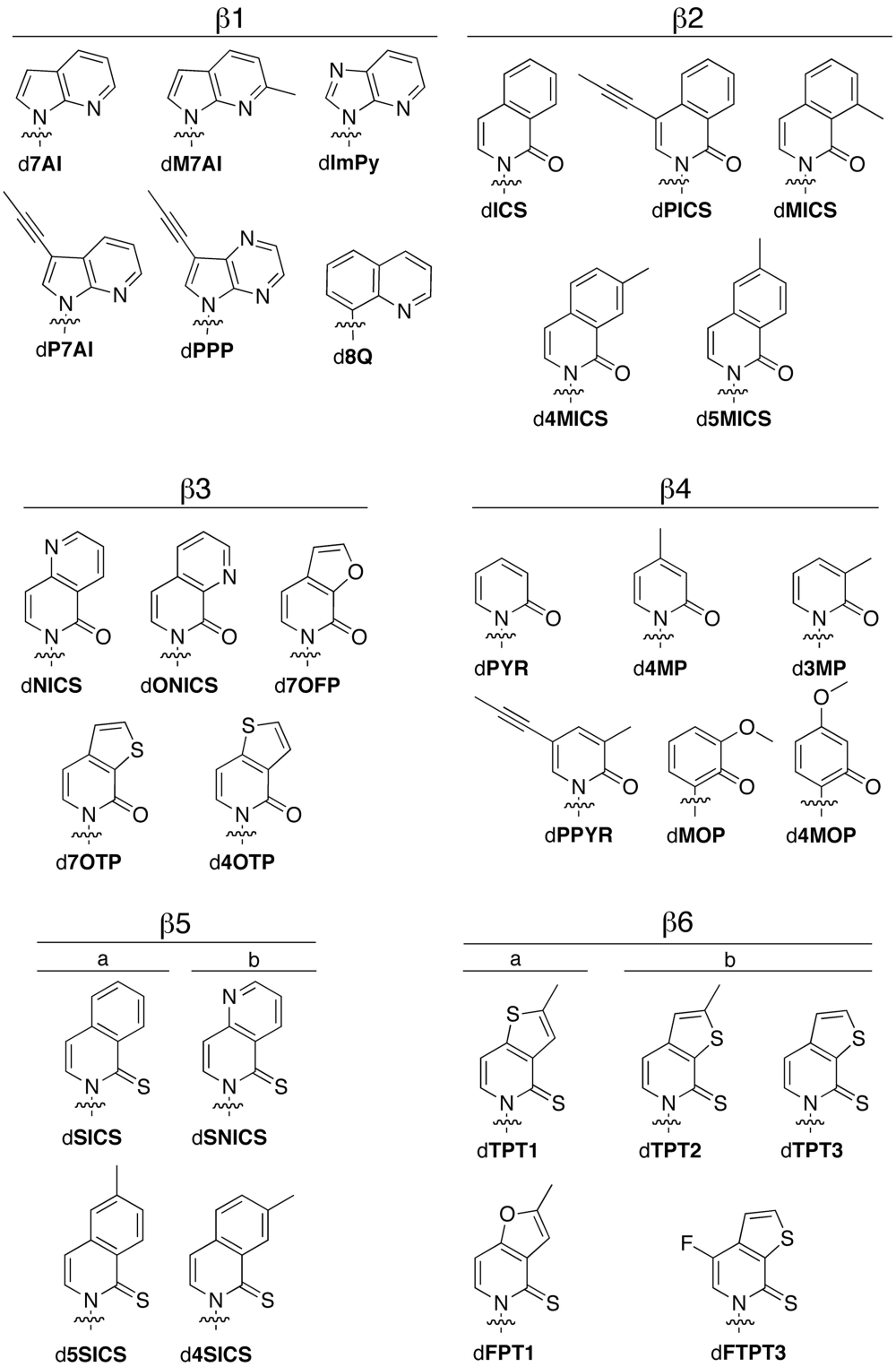

Figure 3. β group unnatural deoxynucleoside triphosphates. Sugar and phosphate backbone are omitted for clarity. References for each compound are provided in Supplementary Table S6.

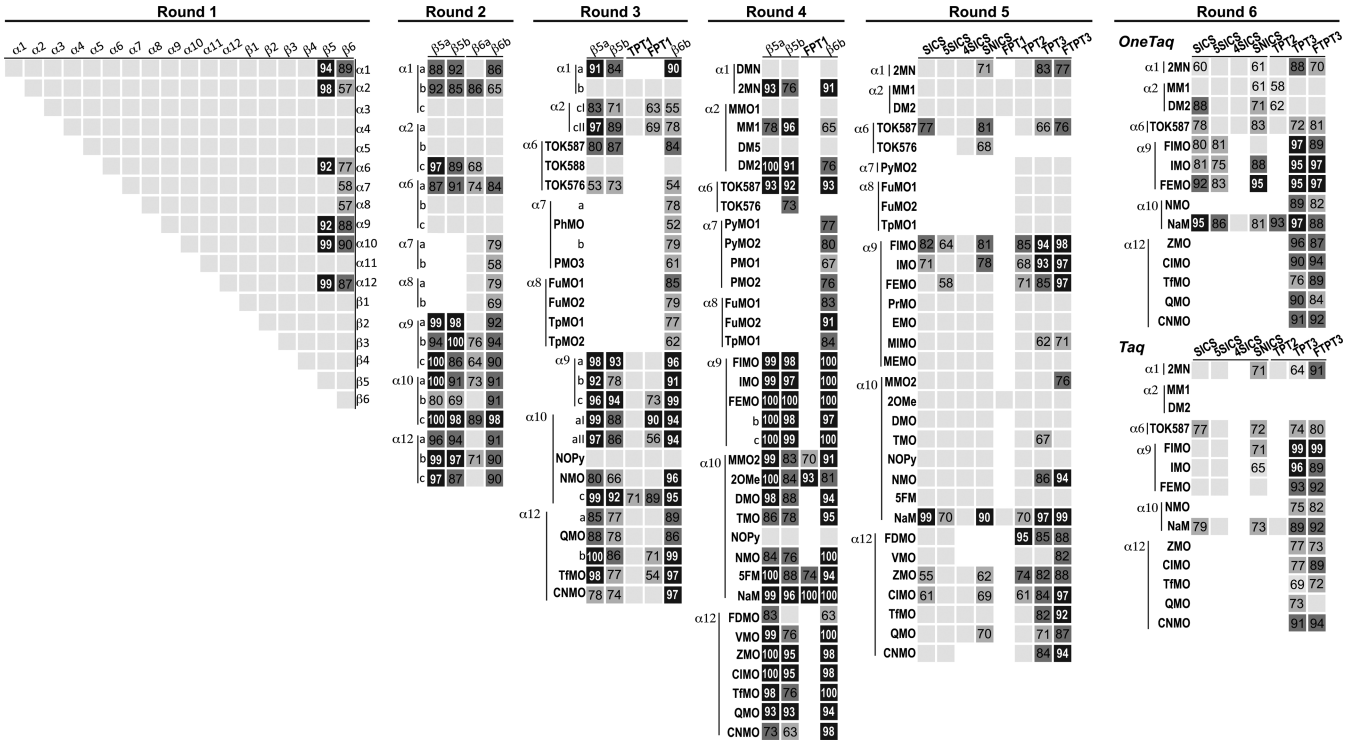

Figure 4. UBP retention (%) after PCR amplification during each round of screening. Open squares indicate UBPs that were not evaluated; light gray squares indicate UPBs that were replicated with less than 50% retention, while those that resulted in higher retention are indicated with darker shading and with the retention value included.

Table 1. Characterization of the most promising UBPs

dβTP	dαTP	Amplification, $\times 10^{12}$	Retention, %	Fidelity per doubling, %
			Taq, 10 s Extension	
TPT3	**FIMO**	8.5	84 ± 3	99.60 ± 0.09
	IMO	6.3	81 ± 5	99.50 ± 0.15
	FEMO	5.0	79 ± 3	99.44 ± 0.09
	NaM	5.8	86.5 ± 0.5	99.66 ± 0.01
FTPT3	**FIMO**	4.8	84 ± 3	99.60 ± 0.09
	IMO	5.6	82 ± 5	99.54 ± 0.13
	FEMO	5.7	81 ± 4	99.51 ± 0.11
	NaM	3.7	91 ± 6	99.76 ± 0.15
5SICS	**NaM**	9.3	<50[b]	<85[b]
			OneTaq, 1 min Extension	
TPT3	**FIMO**	8.7	84.7 ± 1.1	99.61 ± 0.03
	IMO	9.4	82.9 ± 1.7	99.56 ± 0.05
	FEMO	10.4	82.2 ± 1.0	99.55 ± 0.03
	NaM	8.3	91.2 ± 1.3	99.79 ± 0.03
FTPT3	**FIMO**	8.2	86 ± 3	99.65 ± 0.08
	IMO	7.1	76.8 ± 1.6	99.38 ± 0.05
	FEMO	6.3	72.4 ± 1.4	99.24 ± 0.04
	NaM	7.0	90 ± 2	99.76 ± 0.06
5SICS	**NaM**	8.1	77.1 ± 0.7	99.00 ± 0.02

[a]Retention and fidelity determined as described in Materials and Methods.
[b]UBP retention below 50%, and fidelity is thus estimated to be <85%.

nucleotides. Moreover, the denaturation and annealing steps were decreased to 5 s each, and the extension time was decreased to 10 s. Under these conditions, we explored amplification either with OneTaq or with Taq alone. The results with OneTaq showed the highest retention (>95%) with d**SICS**TP and d**NaM**TP; d**SNICS**TP and d**FEMO**TP; d**TPT3**TP and d**FIMO**TP, d**IMO**TP, d**FEMO**TP, d**ZMO**TP or d**NaM**TP and d**FTPT3**TP and d**IMO**TP or d**FEMO**TP. Moderate retention (86–94%) was observed with d**SICS**TP and d**FEMO**TP or d**DM2**TP; d**5SICS**TP and d**NaM**TP; d**SNICS**TP or d**IMO**TP; d**TPT2**TP and d**NaM**TP; d**TPT3**TP and d**NMO**TP, d**ClMO**TP, d**QMO**TP, d**CNMO**TP or d**2MN**TP and d**FTPT3**TP and d**FIMO**TP, d**NaM**TP,

Table 2. Characterization of additional promising UBPs

dβTP	dαTP	Retention (%)
SICS	**NaM**	99[a]
SICS	**FEMO**	92[b]
SNICS	**NaM**	90[a]
SNICS	**FEMO**	95[b]
SNICS	**IMO**	88[b]
TPT3	**NMO**	89[b]
TPT3	**ZMO**	86[b]
TPT3	**ClMO**	90[b]
TPT3	**QMO**	90[b]
TPT3	**CNMO**	91[b]
FTPT3	**NMO**	94[a]
FTPT3	**ZMO**	88[a]
FTPT3	**ClMO**	97[a]
FTPT3	**QMO**	87[a]
FTPT3	**CNMO**	94[a]

[a]PCR Conditions: 100 pg D8 template (2) amplified for 16 cycles with Taq polymerase under thermocycling conditions: initial denaturation at 96°C for 1 min, 96°C for 30 s, 60°C for 15 s, 68°C for 60 s.
[b]PCR Conditions: 10 pg D6 template (26) amplified for 24 cycles with OneTaq polymerase under thermocycling conditions: initial denaturation at 96°C for 1 min, 96°C for 5 s, 60°C for 5 s, 68°C for 10 s.

d**ZMO**TP, d**ClMO**TP, d**TfMO**TP or d**CNMO**TP. While retention during Taq-mediated amplification was in general reduced relative to that with OneTaq, the general trends were similar. The highest retention (>96%) was observed with d**TPT3**TP and d**FIMO**TP or d**IMO**TP, and with d**FTPT3**TP and d**FIMO**TP. Only slightly lower retention (89–94%) was observed with d**TPT3**TP and d**FEMO**TP, d**NaM**TP or d**CNMO**TP and d**FTPT3**TP and d**IMO**TP, d**FEMO**TP, d**NaM**TP, d**ClMO**TP, d**CNMO**TP or d**2MN**TP.

Amplification with the most promising combinations of triphosphates, d**TPT3**TP or d**FTPT3**TP and d**FIMO**TP, d**IMO**TP, d**FEMO**TP or d**NaM**TP, was then performed over 52 cycles with Taq and a 10 s extension time, to explore particularly stringent conditions, or with OneTaq and a 30 s extension time, to explore more practical conditions (Table 1, Figure 5). Both amplified strands were sequenced in triplicate to determine UBP retention with high accuracy. With Taq, d**TPT3**-d**NaM**, d**TPT3**-d**FIMO**, d**FTPT3**-d**NaM** and d**FTPT3**-d**FIMO** showed the highest retention, while the pairs involving d**IMO** and d**FEMO** showed somewhat less retention. With OneTaq, d**TPT3**-d**NaM** and d**FTPT3**-d**NaM** showed the highest retention, followed closely by d**FTPT3**-d**FIMO** and d**TPT3**-d**FIMO**.

The screening data suggest that several pairs formed between d**TPT3** and the previously unexamined pyridine-based derivatives of α6 were reasonably well replicated. Thus, we examined in triplicate the amplification of DNA containing these UBPs using OneTaq and 16 amplification cycles with 1 min extension times (Supplementary Table S4). The pairs formed between d**TPT3** and d**TOK580**, d**TOK582** or d**TOK586** were poorly replicated. However, the pairs formed between d**TPT3** and d**TOK588**, d**TOK581**, d**TOK576** and d**TOK587** were amplified with a retention of 62%, 65%, 85% and 94%, respectively.

Finally, the screening data suggested that the pairs formed between d**TPT3** and d**2MN** or d**DM2** are reasonably well replicated, despite neither d**2MN** nor d**DM2** possessing a putatively essential *ortho* H-bond acceptor. Thus, these pairs were further examined via 16 cycles of amplification with OneTaq or Taq alone, and with extension times of either 1 min or 10 s (Supplementary Table S5). With Taq alone, only poor retention was observed. However, with OneTaq, retention was better for both pairs. Retention of the d**TPT3**-d**DM2** pair is 58% and 69% with 1 min and 10 s extension times, respectively. Remarkably, d**TPT3**-d**2MN** is amplified with retentions of 96% and 94% with 1 min and 10 s extension times, respectively.

DISCUSSION

By relying on the propagation of DNA containing d**5SICS**-d**NaM**, we have recently succeeded in creating the first semi-synthetic organism with an expanded genetic alphabet (20). Nonetheless, the creation of semi-synthetic organisms that indefinitely retain the UBP in all possible sequence contexts, including those that are difficult to replicate or that contain multiple UBPs, will likely be facilitated by the availability of multiple, structurally distinct UBPs. Significant progress toward this goal was recently reported with discovery of the d**TPT3**-d**NaM** UBP (23). However, as we have synthesized more analogs of d**5SICS**/d**TPT3** (referred to herein as β derivatives) and d**NaM** (α derivatives) than can be evaluated individually, it was not clear whether d**TPT3**-d**NaM** was even the best UBP among those already available. Thus, we initiated a PCR-based screen to identify the most promising UBPs. In addition, to increase the SAR content of the screen, we included seven novel α derivatives that are based on a pyridyl scaffold with different substituents at the positions *ortho* and *para* to the glycosidic linkage.

SAR data

Even under permissive conditions, where exonucleotidic proofreading activity was present and extension times were 1 min, only mixed groupings of α analogs with β analogs showed significant levels of retention, demonstrating that efficient replication requires the pairing of an α scaffold with a β scaffold. However, the only dβ groups that showed high

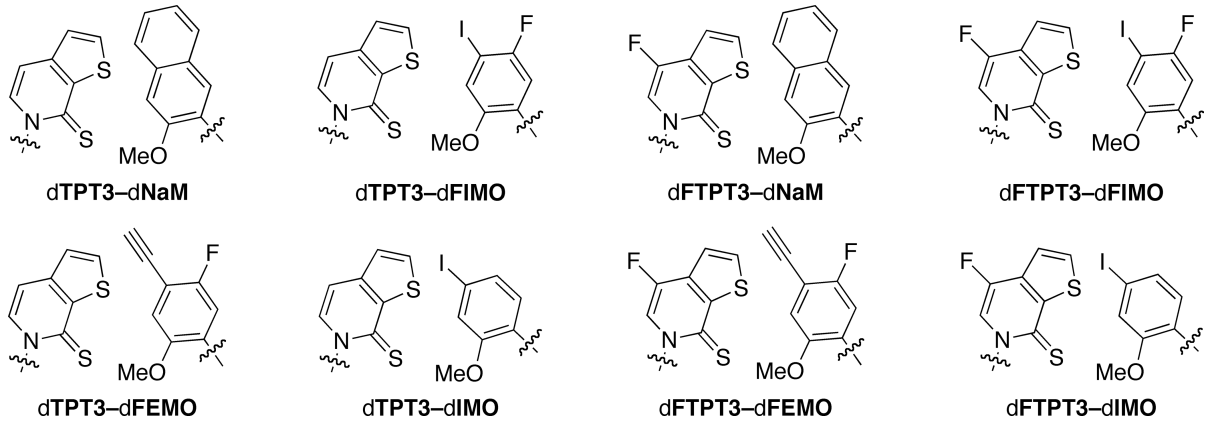

Figure 5. UBPs identified by the present study.

retention were β5 and β6. This reveals the privileged status of the d**5SICS**/d**TPT3**-like scaffold relative to all of the others examined. Surprisingly, the dominant contribution to the high retention with group β5 proved to result not from pairs involving d**5SICS**, but rather from pairs involving d**SICS**, and to a lesser extent d**SNICS**. For example, under all conditions, d**SICS**-d**NaM** was better replicated than d**5SICS**-d**NaM**. Interestingly, d**5SICS** resulted from the optimization of d**SICS** for pairing with d**MMO2** (10); apparently, the increased bulk of d**NaM** makes the added methyl group deleterious. Furthermore, d**SNICS**-d**NaM** is replicated nearly as well (with OneTaq) or better (with Taq) than d**5SICS**-d**NaM**, suggesting that a 6-aza substituent optimizes UBP synthesis by facilitating insertion of the unnatural triphosphate opposite d**NaM** or by increasing the efficiency with which the unnatural nucleotide templates the insertion of d**NaM**TP. Finally, d**SNICS**-d**FEMO** is also better replicated than d**5SICS**-d**NaM**, but only in the presence of proofreading, suggesting that while triphosphate insertion may be less efficient, increased efficiency of extension results in an overall increase in fidelity. The dominant contribution to high-fidelity retention with group β6 was provided by d**TPT3** and d**FTPT3**. In general, both paired well with d**NaM**, d**FEMO**, d**FIMO** or d**IMO**. d**TPT3** paired especially well with d**FIMO** and d**IMO**, suggesting that the *para* iodo substituent mediates favorable interactions, and it also paired well with d**FEMO** and especially d**NaM** when exonuclease activity was present. d**FTPT3** paired well with either d**IMO** or d**FEMO** in the presence of exonuclease activity, as well as with d**FIMO** and d**NaM** in its absence.

While the nitrogen substituent of the pyridine-based α analogs (group α6) was generally detrimental for replication, a more detailed analysis of the UBPs formed with d**TPT3** revealed several interesting trends. As has been observed with other scaffolds, a methyl, chloro or amino substituent at the position *ortho* to the *C*-glycosidic linkage resulted in poorly replicated pairs, presumably due to poor extension after incorporation of the unnatural triphosphate. Also as has been observed with other scaffolds, the *ortho* methoxy substituent of d**TOK581** resulted in better replication, presumably due to its ability to both hydrophobically pack with the template during UBP synthesis and accept an H-bond with a polymerase-based H-bond donor during extension (10). Surprisingly, the data also revealed that the methylsulfanyl *ortho* substituent of d**TOK588**, d**TOK576** and especially d**TOK587** results in better replication. This improvement is likely due to a more optimized compromise between the ability to hydrophobically pack and the ability to accept an H-bond from the polymerase at the primer terminus. We also found that the *para* substituent in this series of derivatives can contribute to efficient replication, with a bromo substituent being the best, followed by a second methylsulfanyl group, and then finally a simple methyl group. When d**TOK587**, with its combination of the *ortho* methylsulfanyl and *para* bromo substituents, was paired with d**TPT3**, the resulting UBP was replicated by OneTaq and 1 min extension times with a fidelity (calculated from retention level as reported previously (2,26)) of 99.3%, which is slightly better than d**5SICS**-d**MMO2** under similar conditions. Clearly, similar *ortho* methylsulfanyl and *para* bromo substituents should be examined with the more efficiently replicated α-like scaffolds, such as d**FIMO** and d**NaM**.

The replication of the pairs formed between d**TPT3** and d**2MN** or d**DM2** also merits discussion. DNA containing these pairs is not amplified by Taq alone, but is surprisingly well amplified by OneTaq. This result was unexpected because neither d**2MN** nor d**DM2** possesses the *ortho* H-bond acceptor that has been postulated to be essential for extension of the nascent (natural or unnatural) primer terminus. Specifically, when a nucleotide is positioned at the growing primer terminus, the H-bond acceptor is disposed into the developing minor groove where it accepts an H-bond from the polymerase, and this H-bond is thought to be required for proper terminus alignment (14–16). When amplified with OneTaq and a 1 min extension time, d**TPT3**-d**2MN** is replicated with a fidelity of 99.5%, which only drops to 99.1% when the extension time is reduced to 10 s. The absence of amplification in the absence of proofreading, coupled with the only small decrease observed in the presence of proofreading when extension times were reduced, implies that the surprisingly high-fidelity amplification of DNA containing d**TPT3**-d**2MN** results from selective extension of the UBP relative to mispairs. This suggests that the absence of an *ortho* H-bond acceptor is more dele-

terious for the extension of a mispair than for the extension of the UBPs.

Efforts toward the expansion of the genetic alphabet

Overall, the data confirms that d**TPT3**-d**NaM** is the most promising UBP of those currently available, and current efforts toward the expansion of the genetic code will focus on this UBP. However, the pairs formed between d**TPT3** and d**FEMO**, d**FIMO** or d**IMO**, or between d**FTPT3** and d**NaM**, d**FEMO**, d**FIMO** or d**IMO**, are also particularly promising. Given that each of these eight pairs is replicated more efficiently than d**5SICS**-d**NaM**, and that d**5SICS**-d**NaM** is sufficiently well replicated to be stably propagated within a cell (20), each of these UBPs is a viable candidate for use in the expansion of an organism's genetic code. Clearly, the core scaffolds represented by d**TPT3** and d**NaM** are a general solution to the challenge of storing genetic information, a property previously only associated with the purines and pyrimidines of the natural nucleotides.

In addition to the most promising UBPs noted above, it is noteworthy that a remarkable number of additional novel pairs are replicated with only a moderately reduced fidelity, or are replicated with a high fidelity when the amplification is performed under less stringent conditions (Table 2). Along with the most efficiently replicated UBPs, these pairs provide a wide range of scaffolds with diverse physicochemical properties for further optimization efforts. This is especially critical in the effort to optimize *in vivo* replication, where different physicochemical properties are expected to bestow the constituent nucleotides with different pharmacokinetic-like properties, the optimization of which is also likely to be important during the effort to create stable and healthy semi-synthetic organisms that are able to store and retrieve increased genetic information.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported by the National Institutes of Health [GM060005 to F.E.R]; Czech Science Foundation [P206/12/G151 to M.H.]. Source of open access funding: National Institutes of Health.

Conflict of interest statement. F.E.R. and D.A.M. have a financial interest (shares) in Synthorx Inc., a company that has commercial interest in the UBPs.

REFERENCES

1. McMinn, D.L., Ogawa, A.K., Wu, Y., Liu, J., Schultz, P.G., and Romesberg, F.E. McMinn, D.L., Ogawa, A.K., Wu, Y., Liu, J., Schultz, P.G., and Romesberg, F.E. (1999) Efforts toward expansion of the genetic alphabet: DNA polymerase recognition of a highly stable, self-pairing hydrophobic base. *J. Am. Chem. Soc.*, **121**, 11585–11586.
2. Malyshev, D.A., Dhami, K., Quach, H.T., Lavergne, T., Ordoukhanian, P., Torkamani, A., and Romesberg, F.E.Malyshev, D.A., Dhami, K., Quach, H.T., Lavergne, T., Ordoukhanian, P., Torkamani, A., and Romesberg, F.E. (2012) 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.*, **109**, 12005–12010.
3. Betz, K., Malyshev, D.A., Lavergne, T., Welte, W., Diederichs, K., Dwyer, T.J., Ordoukhanian, P., Romesberg, F.E., and Marx, A.Betz, K., Malyshev, D.A., Lavergne, T., Welte, W., Diederichs, K., Dwyer, T.J., Ordoukhanian, P., Romesberg, F.E., and Marx, A. (2012) KlenTaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. *Nat. Chem. Biol.*, **8**, 612–614.
4. Seo, Y.J., Malyshev, D.A., Lavergne, T., Ordoukhanian, P., and Romesberg, F.E.Seo, Y.J., Malyshev, D.A., Lavergne, T., Ordoukhanian, P., and Romesberg, F.E. (2011) Site-specific labeling of DNA and RNA using an efficiently replicated and transcribed class of unnatural base pairs. *J. Am. Chem. Soc.*, **133**, 19878–19888.
5. Yang, Z., Chen, F., Alvarado, J.B., and Benner, S.A.Yang, Z., Chen, F., Alvarado, J.B., and Benner, S.A. (2011) Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *J. Am. Chem. Soc.*, **133**, 15105–15112.
6. Kaul, C., Muller, M., Wagner, M., Schneider, S., and Carell, T.Kaul, C., Muller, M., Wagner, M., Schneider, S., and Carell, T. (2011) Reversible bond formation enables the replication and amplification of a crosslinking salen complex as an orthogonal base pair. *Nat. Chem.*, **3**, 794–800.
7. Kimoto, M., Kawai, R., Mitsui, T., Yokoyama, S., and Hirao, I.Kimoto, M., Kawai, R., Mitsui, T., Yokoyama, S., and Hirao, I. (2009) An unnatural base pair system for efficient PCR amplification and functionalization of DNA molecules. *Nucleic Acids Res.*, **37**, e14.
8. Switzer, C., Moroney, S.E., and Benner, S.A.Switzer, C., Moroney, S.E., and Benner, S.A. (1989) Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.*, **111**, 8322–8323.
9. Henry, A.A. and Romesberg, F.E.Henry, A.A. and Romesberg, F.E. (2003) Beyond A, C, G and T: augmenting nature's alphabet. *Curr. Opin. Chem. Biol.*, **7**, 727–733.
10. Leconte, A.M., Hwang, G.T., Matsuda, S., Capek, P., Hari, Y., and Romesberg, F.E. Leconte, A.M., Hwang, G.T., Matsuda, S., Capek, P., Hari, Y., and Romesberg, F.E. (2008) Discovery, characterization, and optimization of an unnatural base pair for expansion of the genetic alphabet. *J. Am. Chem. Soc.*, **130**, 2336–2343.
11. Hwang, G.T. and Romesberg, F.E.Hwang, G.T. and Romesberg, F.E. (2008) Unnatural substrate repertoire of A, B, and X family DNA polymerases. *J. Am. Chem. Soc.*, **130**, 14872–14882.
12. Matsuda, S., Leconte, A.M., and Romesberg, F.E.Matsuda, S., Leconte, A.M., and Romesberg, F.E. (2007) Minor groove hydrogen bonds and the replication of unnatural base pairs. *J. Am. Chem. Soc.*, **129**, 5551–5557.
13. Yu, C., Henry, A.A., Romesberg, F.E., and Schultz, P.G.Yu, C., Henry, A.A., Romesberg, F.E., and Schultz, P.G. (2002) Polymerase recognition of unnatural base pairs. *Angew. Chem. Int. Ed.*, **41**, 3841–3844.
14. Meyer, A.S., Blandino, M., and Spratt, T.E.Meyer, A.S., Blandino, M., and Spratt, T.E. (2004) *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.*, **279**, 33043–33046.
15. Morales, J.C. and Kool, E.T.Morales, J.C. and Kool, E.T. (1999) Minor groove interactions between polymerase and DNA: more essential to replication that Watson-Crick hydrogen bonds. *J. Am. Chem. Soc.*, **121**, 2323–2324.
16. Spratt, T.E.Spratt, T.E. (2001) 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*, **40**, 2647–2652.
17. Lavergne, T., Malyshev, D.A., and Romesberg, F.E.Lavergne, T., Malyshev, D.A., and Romesberg, F.E. (2012) Major groove substituents and polymerase recognition of a class of predominantly hydrophobic unnatural base pairs. *Chem. Eur. J.*, **18**, 1231–1239.
18. Seo, Y.J., Matsuda, S., and Romesberg, F.E.Seo, Y.J., Matsuda, S., and Romesberg, F.E. (2009) Transcription of an expanded genetic alphabet. *J. Am. Chem. Soc.*, **131**, 5046–5047.
19. Li, Z., Lavergne, T., Malyshev, D.A., Zimmermann, J., Adhikary, R., Dhami, K., Ordoukhanian, P., Sun, Z., Xiang, J., and Romesberg, F.E.Li, Z., Lavergne, T., Malyshev, D.A., Zimmermann, J., Adhikary, R., Dhami, K., Ordoukhanian, P., Sun, Z., Xiang, J., and Romesberg, F.E. (2013) Site-specifically arraying small molecules or proteins on

DNA using an expanded genetic alphabet. *Chem. Eur. J.*, **19**, 14205–14209.
20. Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa, I.R.J., and Romesberg, F.E.Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa, I.R.J., and Romesberg, F.E. (2014) A semi-synthetic organism with an expanded genetic alphabet. *Nature*, **509**, 385–388.
21. Betz, K., Malyshev, D.A., Lavergne, T., Welte, W., Diederichs, K., Romesberg, F.E., and Marx, A.Betz, K., Malyshev, D.A., Lavergne, T., Welte, W., Diederichs, K., Romesberg, F.E., and Marx, A. (2013) Structural insights into DNA replication without hydrogen bonds. *J. Am. Chem. Soc.*, **135**, 18637–18643.
22. Lavergne, T., Degardin, M., Malyshev, D.A., Quach, H.T., Dhami, K., Ordoukhanian, P., and Romesberg, F.E.Lavergne, T., Degardin, M., Malyshev, D.A., Quach, H.T., Dhami, K., Ordoukhanian, P., and Romesberg, F.E. (2013) Expanding the scope of replicable unnatural DNA: stepwise optimization of a predominantly hydrophobic base pair. *J. Am. Chem. Soc.*, **135**, 5408–5419.
23. Li, L., Degardin, M., Lavergne, T., Malyshev, D.A., Dhami, K., Ordoukhanian, P., and Romesberg, F.E.Li, L., Degardin, M., Lavergne, T., Malyshev, D.A., Dhami, K., Ordoukhanian, P., and Romesberg, F.E. (2014) Natural-like replication of an unnatural base pair for the expansion of the genetic alphabet and biotechnology applications. *J. Am. Chem. Soc.*, **136**, 826–829.
24. Kubelka, T., Slavetinska, L., Eigner, V., and Hocek, M.Kubelka, T., Slavetinska, L., Eigner, V., and Hocek, M. (2013) Synthesis of 2,6-disubstituted pyridin-3-yl C-2'-deoxyribonucleosides through chemoselective transformations of bromo-chloropyridine C-nucleosides. *Org. Biomol. Chem.*, **11**, 4702–4718.
25. Ludwig, J. and Eckstein, F.Ludwig, J. and Eckstein, F. (1989) Rapid and efficient synthesis of nucleoside 5'-0-(1-thiotriphosphates), 5'-triphosphates and 2',3'-cyclophosphorothioates using 2-chloro-4H-1,3,2-benzodioxaphosphorin-4-one. *J. Org. Chem.*, **54**, 631–635.
26. Malyshev, D.A., Seo, Y.J., Ordoukhanian, P., and Romesberg, F.E.Malyshev, D.A., Seo, Y.J., Ordoukhanian, P., and Romesberg, F.E. (2009) PCR with an expanded genetic alphabet. *J. Am. Chem. Soc.*, **131**, 14620–14621.
27. Filee, J., Forterre, P., Sen-Lin, T., and Laurent, J.Filee, J., Forterre, P., Sen-Lin, T., and Laurent, J. (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J. Mol. Evol.*, **54**, 763–773.
28. Kornberg, A. and Baker, T.A.Kornberg, A. and Baker, T.A. (2005) In: *DNA Replication*, University Science Books, Sausalito, CA..