Published in final edited form as:

J Am Chem Soc. 2019 July 10; 141(27): 10644–10653. doi:10.1021/jacs.9b02075.

Optimization of Replication, Transcription, and Translation in a Semi-Synthetic Organism

Aaron W. Feldman^{†,‡}, Vivian T. Dien^{†,‡}, Rebekah J. Karadeema[‡], Emil C. Fischer[‡], Yanbo You[§], Brooke A. Anderson[‡], Ramanarayanan Krishnamurthy[‡], Jason S. Chen[‡], Lingjun Li^{§,*}, Floyd E. Romesberg^{‡,*}

[‡]Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

[§]School of Chemistry and Chemical Engineering, Henan Normal University, Henan 453007, P. R. China

Abstract

Previously, we reported the creation of a semi-synthetic organism (SSO) that stores and retrieves increased information by virtue of stably maintaining an unnatural base pair (UBP) in its DNA, transcribing the corresponding unnatural nucleotides into the codons and anticodons of mRNAs and tRNAs, and then using them to produce proteins containing non-canonical amino acids (ncAAs). Here we report a systematic extension of the effort to optimize the SSO by exploring a variety of deoxy- and ribonucleotide analogs. Importantly, this includes the first *in vivo* structure-activity relationship (SAR) analysis of unnatural ribonucleoside triphosphates. Similarities and differences between how DNA and RNA polymerases recognize the unnatural nucleotides were observed, and remarkably, we found that a wide variety of unnatural ribonucleotides can be efficiently transcribed into RNA and then productively and selectively paired at the ribosome to mediate the synthesis of proteins with ncAAs. The results extend previous studies, demonstrating that nucleotides bearing no significant structural or functional homology to the natural nucleotides can be efficiently and selectively paired during replication, to include each step of the entire process of information storage and retrieval. From a practical perspective, the results identify the most optimal UBP for information storage, as well as the most optimal unnatural ribonucleoside triphosphates for its retrieval. The optimized SSO is now, for the first time, able to efficiently produce proteins containing multiple, proximal ncAAs.

Graphical Abstract

*Corresponding Author floyd@scripps.edu, lingjunlee@htu.edu.cn.

[†]These authors contributed equally.

Author Contributions

The manuscript was written through contributions of all authors.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Methods, Schemes S1–S12, Tables S1–S4, Figures S1–S3, plasmid sequences (PDF)

Notes

The authors declare the following competing financial interests: a patent application has been filed based on the use of UBPs in SSOs and F.E.R. has a financial interest (shares) in Synthorx, Inc., a company that has commercial interests in the UBP.

INTRODUCTION

In all natural organisms, information is encoded with the four-letter genetic alphabet, consisting of deoxyadenosine (dA), deoxyguanosine (dG), deoxycytidine (dC), and deoxythymidine (dT), with the storage and retrieval of this information made possible by the formation of two base pairs, (d)A-dT/U and (d)G-(d)C (where “(d)” indicates that the nucleotides may be either deoxyribo- or ribonucleotides). Over 100 years ago, the newly defined field of synthetic biology set its central goal as the creation of new forms and functions,¹ and perhaps the most general route to this goal is to increase the information that a cell can store and retrieve. Correspondingly, the last decade has seen significant effort and progress toward the identification of a fifth and sixth nucleotide that pair to form a third, unnatural base pair (UBP).^{2–4}

Our efforts have focused on the development of UBPs formed between synthetic nucleotide analogs with predominantly hydrophobic nucleobases that pair via hydrophobic and packing forces, as opposed to the complementary hydrogen bonding used by natural nucleobases. Through a medicinal chemistry-like approach, based on the elucidation of structure-activity relationships (SARs) from the evaluation of over 150 nucleotide analogs,⁵ we discovered a family of UBPs that, when incorporated into DNA, are well replicated by DNA polymerases *in vitro*. Within this family, the UBPs formed between the synthetic nucleotides dNaM and either d5SICS or dTPT3 (dNaM-d5SICS and dNaM-dTPT3, respectively; Figure 1) have received the most attention and are amongst the most promising. Towards the goal of creating new forms and functions, we have used these UBPs as the basis of a semi-synthetic organism (SSO). The SSO, a strain of *Escherichia coli* constitutively expressing a nucleoside triphosphate transporter from *Phaedactylum tricornutum* (PNTT2),⁶ imports the requisite unnatural nucleoside triphosphates from the media and uses them to replicate DNA containing the UBP, transcribe mRNA and tRNA containing the unnatural nucleotides, and then use the resulting cognate unnatural codons and anticodons to translate proteins containing non-canonical amino acids (ncAAs).^{7–12} In addition, the forces underlying the pairing of the unnatural nucleotides, as well as their physical properties have been explored by others.^{13–19}

Although dNaM-d5SICS, and especially dNaM-dTPT3, are sufficiently well replicated in the SSO for the stable propagation of increased genetic information, they are still replicated less efficiently than a natural base pair, which has motivated continued chemical optimization. These efforts have identified several promising candidates, in particular

dCNMO and dPTMO, which when paired with dTPT3, form UBPs that are better retained in the DNA of the SSO (Figure 1).^{10,20} In addition, we have also explored the ability of the SSO to retrieve the information encoded with the dPTMO-dTPT3 UBP via transcription and translation, and interestingly, found that its use results in more efficient expression of unnatural proteins than dNaM-dTPT3.¹⁰ This result clearly motivates elucidation of the SARs governing the templating of transcription. In addition, only the NaMTP and TPT3TP ribonucleoside triphosphates have been used to retrieve the increased information in the SSO. Given that these analogs were designed based on replication SARs, which may or may not be the same as those governing efficient transcription and translation, it remains unclear if they are optimal.

Here, we explore the ability of the unnatural information encoded by UBPs formed by dTPT3 and either dNaM or a dNaM analog, to be retrieved in the form of proteins containing ncAAs. We also report the first systematic evaluation of the efficiency with which this information is retrieved using thirteen analogs of NaMTP or TPT3TP, eleven of which are novel. The results identify an improved system of unnatural deoxy- and ribonucleotides that form the basis of an SSO that more efficiently stores and retrieves increased information, and for the first time, permits the efficient production of protein bearing multiple, proximal ncAAs.

RESULTS

Replication and templating of transcription.

While we have examined the *in vivo* replication of a wide variety of our UBPs, retention in a tRNA gene or in any actively transcribed gene has to date only been examined with dNaM, dPTMO, dMTMO, and dTPT3.^{9–10} To extend these studies, we first constructed plasmids with two dNaM-dTPT3 UBPs, such that the sequence AXC (here and throughout **X** refers to (d)NaM or a (d)NaM analog) was positioned to template codon 151 of sfGFP mRNA (sfGFP¹⁵¹(AXC)) and with the sequence GYT (here and throughout **Y** refers to (d)TPT3 or a (d)TPT3 analog) positioned to template the anticodon of the *M. mazei* Pyl tRNA (tRNA^{Pyl}(GYT)), which is selectively charged with the ncAA *N*⁶-(2-azidoethoxy)-carbonyl-L-lysine (AzK) by the *M. barkeri* pyrrolysyl-tRNA synthetase (*Mb* PylRS).^{21–25} These plasmids were used to transform *E. coli* expressing the nucleoside triphosphate transporter PNTT2 (strain YZ3⁸) and harboring a plasmid encoding *Mb* PylRS. After transformation, colonies were selected and grown to an OD₆₀₀ ~ 1.0 in liquid media supplemented with dTPT3TP (10 μM) and one of seven different dXTPs (Figure 2) added at varying concentrations (150 μM, 10 μM, or 5 μM). Cells were then diluted into fresh expression media containing the same unnatural deoxyribonucleotides as well as NaMTP (250 μM), TPT3TP (30 μM), and AzK (10 mM). After a brief incubation, T7 RNAP and tRNA^{Pyl}(GYT) expression was induced by the addition of isopropyl-β-d-thiogalactoside (IPTG, 1 mM). After an additional 1 h of incubation, the expression of sfGFP¹⁵¹(AXC) was initiated by the addition of anhydrotetracycline (aTc, 100 ng/mL).

After 2.5 h, plasmids were isolated and then the genes of interest were independently PCR amplified using d5SICSTP and dMMO2^{BIO}TP, a biotinylated analog of dNaMTP.²⁶ The

resulting PCR product was analyzed via a gel mobility shift assay using streptavidin to quantify the UBP retained as a percent shift of the total amplified product (hereafter referred to as the streptavidin gel shift assay) (Figure 3A and 3B). At the highest dXTP concentration examined (150 μ M), retention in the sfGFP¹⁵¹(AXC) gene was similar for each dXTP analog, varying from 99% for dNaMTP to 92% for dMTMOTP. Retention within the tRNA^{Pyl}(GYT) gene was slightly lower, ranging from 82% for dMTMOTP to 74% for dMMO2TP. With the addition of 10 μ M of each dXTP, retention in the sfGFP¹⁵¹(AXC) gene ranged from 96% for d5FMTP to 73% for dNaMTP. With the tRNA^{Pyl}(GYT) gene, retentions were again generally slightly lower, ranging from 82% for d5FMTP and dPTMOTP to 71% for dNaMTP. At the lowest concentration (5 μ M), retention ranged from 94% for d5FMTP to 64% for dMTMOTP in the sfGFP¹⁵¹(AXC) gene, and from 83% for d5FMTP to 71% for dMTMOTP in the tRNA^{Pyl}(GYT) gene. When the same concentrations were employed, the results with dNaMTP, dPTMOTP, or dMTMOTP are indistinguishable from those previously reported, and also as reported previously, cells grown with 5 μ M dNaMTP did not survive.¹⁰ Generally, at the highest concentration, all dXTPs performed well with nearly quantitative retention of the UBP within the sfGFP gene. However, at lower concentrations it became increasingly evident that dCNMO-dTPT3, d5FM-dTPT3, and dPTMO-dTPT3 were replicated with significantly higher retention. Interestingly, retention in the tRNA gene was significantly less dependent on dXTP concentration.

To characterize the amount of protein produced, bulk culture fluorescence normalized to cell growth was measured 2.5 h after the induction of protein expression (Figure 3C). In the absence of AzK, fluorescence was generally low, with the exceptions of dPTMOTP and dMTMOTP at the lower concentrations, which appeared slightly higher. When AzK was added to the media, cells grown with each dXTP at either 150 or 10 μ M generally showed significant and similar levels of fluorescence, with the exceptions of dNaMTP, which showed significantly less fluorescence at 10 μ M than at 150 μ M. At the lowest concentration (5 μ M), the addition of AzK resulted in less fluorescence observed with dMTMOTP while it remained the same with dCNMOTP, d5FMTP, dCIMOTP, dMMO2TP, or dPTMOTP. Again, as mentioned above, cells were unable to grow when provided with only 5 μ M dNaMTP. Remarkably, the data indicate that while similar at high concentrations, at lower concentrations the use of each dXTP analog results in a greater AzK-dependent increase in fluorescence than does the use of dNaMTP. In particular, dCNMOTP, d5FMTP, and dCIMOTP showed the greatest AzK-dependent increase in fluorescence, and were thus considered the most promising.

To directly assess the fidelity of unnatural protein production, cells were harvested 2.5 h after the induction of protein expression and the sfGFP produced was purified and subjected to a strain-promoted azide-alkyne cycloaddition reaction²⁷ with dibenzocyclooctyne (DBCO) linked to a TAMRA dye by four PEG units. In addition to tagging the proteins containing the ncAA with a detectable fluorophore, conjugation produces a shift in electrophoretic migration, allowing quantification of protein containing AzK as a percentage of the total protein produced (i.e. fidelity of ncAA incorporation; Figure 3D).¹⁰ As with dNaM, use of each dXTP at the highest concentration (150 μ M) resulted in virtually

complete shifts of the purified protein, reflecting high fidelity incorporation of the ncAA. When grown in the presence of 10 μ M of the dXTP, the fidelity of ncAA incorporation remained high for dCNMOTP, d5FMTP, dCIMOTP, dMMO2TP, dPTMOTP, and dMTMOTP, but dropped precipitously for dNaMTP. Finally, at a concentration of 5 μ M dXTP, the fidelity of ncAA incorporation remained similar to that observed at 10 μ M for dCNMOTP, d5FMTP, dCIMOTP, dMMO2TP, and dPTMOTP, but dropped for dMTMOTP (again due to viability, fidelity with dNaMTP could not be measured at this concentration).

The most promising members of the family that produced the greatest quantity of pure unnatural protein, especially at lower concentrations, are dCNMOTP and d5FMTP. However, relative to d5FMTP, the use dCNMOTP has been previously shown to result in higher UBP retention in more difficult to replicate sequences²⁰. Thus, we conclude that the dCNMO-dTPT3 UBP is the most optimized for the storage and retrieval of increased information.

Synthesis and first SAR analysis of ribonucleotide candidates.

Retrieval of the information made available by the UBP has only previously been explored using NaMTP and TPT3TP. To begin to elucidate the SARs governing efficient transcription and translation in the SSO, we designed and synthesized nine novel NaMTP analogs (Figure 4A) and four novel TPT3TP analogs (Figure 4B). These analogs were designed to explore the role of nucleobase shape, aromatic surface area, and heteroatom derivatization. Generally, the synthesis of the XTP analogs proceeded via lithiation of the corresponding aryl halide, followed by coupling of the lithiated species to either the benzyl- or TBS-protected ribolactone. Reduction of the resulting hemi-acetal intermediate in the presence of boron trifluoride diethyl etherate and triethylsilane afforded the desired protected nucleoside in each case. Following deprotection, the resulting X nucleoside analogs were converted to triphosphates using standard Ludwig phosphorylation conditions (see Supporting Information).²⁸ NaMTP and MMO2TP were synthesized as reported previously.²⁹ The synthesis of the YTP analogs generally proceeded via intramolecular Curtius rearrangement of the corresponding acyl azide, followed by Lewis-acid mediated coupling to 1-*O*-acetyl-2,3,5-tri-*O*-benzoyl- β -D-ribofuranose, resulting in pure β -anomer of the desired protected nucleoside. Following conversion of the pyridine to the corresponding thio-pyridone and subsequent benzoyl deprotection, the corresponding free nucleosides were converted to triphosphates using standard Ludwig phosphorylation conditions (see Supporting Information).²⁸ 5SICSTP was synthesized as reported previously.²⁹

We initiated our SAR analysis with NaMTP and the nine XTP analogs. Based on performance in the dXTP screen described above, and to eliminate variable loss at the DNA level as a complicating factor, we encoded the unnatural information with the dCNMO-dTPT3 UBP. Additionally, we used our recently reported *E. coli* strain ML2,¹¹ which expresses the nucleoside triphosphate transporter P_{NTT}2, but has also been genetically engineered for higher fidelity replication of the UBP by deletion of the gene encoding RecA (as is common in cloning strains) and overexpression of DNA Pol II. The same plasmid described above, harboring sfGFP¹⁵¹(AXC) and tRNA^{Py1}(GYT), was used, and to focus the

screen to a single unnatural ribonucleotide, the *M. mazei* Pyl tRNA was transcribed in the presence of 30 μM **TPT3TP**.

Cells were grown and induced to produce protein as described above, except that the various **XTPs** were provided at either high (250 μM) or low (25 μM) concentration in the expression media (Figure 5). Expressed sfGFP was purified 3 h after induction and ncAA content analyzed using the DBCO-mediated gel shift assay described above (Figure 5B).

Remarkably, the use of each **XTP** at 250 μM resulted in sfGFP gel shifts of at least 63%. Along with the lack of observable shift in the absence of an **XTP**, this demonstrates that each **XTP** is imported by *P_{NTT2}* and participates in transcription and translation at the ribosome with at least reasonable efficiency. While **CNMOTP** and **5F2OMeTP** each performed well, resulting in gel shifts of 92% and 94%, respectively, **NaMTP**, **MMO2TP**, and **5FMTP** performed the best, with protein gel shifts of 98%, 97%, and 98%, respectively. With similarly high levels of protein purity, we were able to compare the relative fluorescence produced using these three **XTPs** in the absence of any complications resulting from significant levels of natural sfGFP contaminant. Cells grown with 250 μM of either **MMO2TP** or **5FMTP** produced 63% and 90% bulk fluorescence, respectively, compared to cells grown with **NaMTP** at the same concentration (Figure 5A).

At the lower concentration tested (25 μM), the use of **NaMTP** resulted in a lower fidelity of ncAA incorporation, with a protein shift that dropped to 86%. While 7 of the 9 **NaMTP** analogs also showed significant decreases in fidelity, the use of **MMO2TP** and **5FMTP** each yielded protein shifts of 94%. Comparing the relative fluorescence of cells grown with these ribonucleotides (again possible due to the similar and high fidelity of unnatural protein produced), **5FMTP** produces 34% more fluorescence than **MMO2TP** at this lower concentration.

We next performed similar experiments, except we supplied a constant amount of **NaMTP** (250 μM) and either a high (250 μM) or low (25 μM) concentration of a **YTP** analog. Consistent with previous reports⁹, the addition of **TPT3TP** at the higher concentration resulted in significantly reduced cell growth and little fluorescence relative to the control sample that did not receive a **YTP** (Figure 5C). In contrast, each of the other **YTP** analogs produced fluorescence above background, and protein shifts of at least 51% (Figure 5D). In particular, **TAT1TP** performed the best at this concentration, with its use resulting in at least 2.6-fold more fluorescence than any other **YTP**, while maintaining a protein shift of 96%. However, the addition of **TAT1TP** at this concentration did result in a modest level of reduced cell growth (Figure S1).

Consistent with previous reports,^{9–10} **TPT3TP** is somewhat less toxic when provided at lower concentrations, and correspondingly, when provided at 25 μM , cells produce significant quantities of pure protein. Under these conditions, **TPT3TP** is more optimal than **SICSTP**, **FSICSTP**, and **5SICSTP**, producing 2-, 5-, and 6-fold greater fluorescence, respectively. Interestingly, the toxicity observed with **TAT1TP** at the higher concentrations was almost completely eliminated at lower concentrations (Figure S1), and its use resulted in even greater fluorescence than at the higher concentration. When provided at 25 μM , **TAT1TP** produces 41% more fluorescence than when provided at 250 μM , and interestingly,

it produced 57% more fluorescence than when **TPT3TP** is provided at 25 μM . Most importantly, the use of **TAT1TP** at these concentrations resulted in the production of protein with 98% ncAA incorporation.

Optimization of unnatural protein production.

The screens described above identified **dCNMO-dTPT3** as the most promising UBP for the storage of information, and **TAT1TP** and **NaMTP** or **5FMTP** as the most promising ribonucleoside triphosphates for its retrieval. Thus, we next turned to exploring the concentrations of each ribonucleotide used to optimize the yield and fidelity of protein expression in the SSO. Upon transformation with the same plasmids used above, 10 μM **dTPT3TP** and 25 μM **dCNMOTP** were provided in the growth media, which was then also supplemented with **TAT1TP** at concentrations ranging from 100 μM to 12.5 μM , and either **NaMTP** or **5FMTP** at concentrations ranging from 200 μM to 12.5 μM , all in series of 2-fold dilutions, and after the addition of 1 mM AzK, the cells were induced to express sfGFP.

Total sfGFP fluorescence observed in cells provided with **TAT1TP** and **NaMTP** was generally higher than that observed in cells provided with **TAT1TP** and **5FMTP** (Figure 6A and 6B). In both cases, fluorescence was higher at lower concentrations of **NaMTP** or **5FMTP** (due to increased production of contaminating natural sfGFP, see below). Additionally, cells generally produced higher fluorescence at higher concentrations of **TAT1TP**. However, due to a slight reduction in growth, cells provided with 100 μM **TAT1TP** produced less fluorescence than those provided with 50 μM **TAT1TP**.

Protein production was again quantified via the gel shift assay (Figure 6C and 6D). Generally, as the concentration of **NaMTP** decreased below 200 μM , incrementally lower fidelity of AzK incorporation into sfGFP was observed, while with use of **5FMTP** this reduction in fidelity was only observed below a concentration of 50 μM . Clearly lower concentrations of **5FMTP** can be used without compromising fidelity. Cells provided with a high concentration of **5FMTP** (50 μM) produced high protein shifts at all concentrations of **TAT1TP** explored (100 μM , 50 μM , 25 μM , or 12.5 μM). However, when the concentration of **5FMTP** was 25 μM or less, decreasing the concentration of **TAT1TP** resulted in a reduced protein shift. When **NaMTP** was provided at 200 μM , all concentrations of **TAT1TP** explored resulted in the production of protein with high fidelity ncAA incorporation, but with lower concentrations of **NaMTP**, decreasing the concentration of **TAT1TP** again resulted in a decreased protein shift.

In all, these studies revealed that the combined optimization of protein purity and yield is achieved with **NaMTP** and **TAT1TP** provided at concentrations of 200 μM and 50 μM , respectively, or with **5FMTP** and **TAT1TP** both provided at a concentration of 50 μM . In terms of protein production alone, the use of **NaMTP** and **TAT1TP** is optimal, whereas the use of **5FMTP** and **TAT1TP** results in slightly lower yields of pure ncAA-labeled protein, but requires significantly lower concentrations of the **XTP**.

Storage and retrieval of higher density unnatural information.

With optimized unnatural nucleotides and conditions, we next sought to challenge the SSO by examining the storage and retrieval of information from a gene containing a higher density of UBPs. Towards this goal, we first validated the ability of the SSO to replicate DNA containing the sfGFP gene with the UBP positioned to encode codons 149 or 153, which are each separated from the codon described above (codon 151) by a single natural codon. Accordingly, expression plasmids were constructed, as described above, but in which the sequence **AXC** was positioned to encode either codon 149 or 153 (sfGFP¹⁴⁹(**AXC**) or sfGFP¹⁵³(**AXC**), respectively). Upon transformation of ML2, cells were grown in the presence of unnatural nucleoside triphosphates, corresponding to either our previously reported system (the deoxyribonucleotides **dNaMTP** and **dTPT3TP** and the ribonucleotides **NaMTP** and **TPT3TP**, denoted **dNaM-dTPT3/NaMTP,TPT3TP**),⁸⁻⁹ or the optimized system discovered in the current study (**dCNMO-dTPT3/NaMTP,TAT1TP**). UBP retention was then characterized using the streptavidin gel shift assay, as described above. High retention of the corresponding UBP in both sfGFP¹⁴⁹(**AXC**) and sfGFP¹⁵³(**AXC**) genes (95%) as well as in the tRNA gene (91%) was observed under both conditions (Table S1).

Total sfGFP fluorescence observed 3 h after induction revealed significant protein production from both constructs in the presence of AzK under both sets of conditions (Figure 7A). However, fluorescence from cells expressing the sfGFP¹⁴⁹(**AXC**) construct provided with **dCNMO-dTPT3/NaM,TAT1** was 58% higher than cells provided with **dNaM-dTPT3/NaM,TPT3**. In the case of the sfGFP¹⁵³(**AXC**) gene, 43% more fluorescence was observed with **dCNMO-dTPT3/NaMTP,TAT1TP** than with **dNaM-dTPT3/NaMTP,TPT3TP**. Under both sets of conditions, approximately 2-fold more fluorescence was observed with sfGFP¹⁵³(**AXC**) than with sfGFP¹⁴⁹(**AXC**). Protein was purified and AzK incorporation was analyzed as described above (Figure 7B). With **dNaM-dTPT3/NaMTP,TPT3TP**, protein shifts of 86% and 94% were observed with sfGFP¹⁴⁹(**AXC**) and sfGFP¹⁵³(**AXC**), respectively. With **dCNMO-dTPT3/NaMTP,TAT1TP**, however, a 96% shift was observed with protein produced from either construct. These results clearly demonstrate that the two additional codon positions are both transcribed and translated efficiently, but again they are transcribed and translated better with the newly identified **dCNMO-dTPT3/NaMTP,TAT1TP** system.

We next constructed expression plasmids with an unnatural codon simultaneously encoded at two or all three of the positions examined (sfGFP^{149,151}(**AXC,AXC**), sfGFP^{151,153}(**AXC,AXC**), sfGFP^{149,153}(**AXC,AXC**), and sfGFP^{149,151,153}(**AXC,AXC,AXC**), respectively). ML2 cells were transformed, grown in the presence of either **dNaM-dTPT3/NaMTP,TPT3TP** or **dCNMO-dTPT3/NaMTP,TAT1TP**, and protein expression was induced as described above. While UBP retention in the tRNA^{Pyl}(**GYT**) gene remained high (88%) in all cases (Table S1), the biotin shift assay with the mRNA genes produced complex and uninterpretable band patterns, likely due to, at least in part, the formation of a mixture of complexes with single PCR products bound to multiple streptavidins. Thus, we proceeded to analyze the protein produced via conjugation to DBCO-TAMRA as described above (Figure 7B). Gratifyingly, relative to the shift observed with a single ncAA, a significantly further shifted band was observed for proteins

expressed from the sfGFP^{149,151}(AXC,AXC), sfGFP^{151,153}(AXC,AXC), and sfGFP^{149,153}(AXC,AXC) constructs, indicating the conjugation of two DBCO-TAMRA molecules to sfGFP bearing two AzK residues. When analyzing purified proteins expressed with dNaM-dTPT3/NaMTP,TPT3TP, quantification of these double shifted bands relative to total sfGFP revealed that 80%, 87%, and 83% of the protein, respectfully, had two AzK residues, and 20%, 13%, or 9% , respectfully, had a single AzK when using the sfGFP^{149,151}(AXC,AXC), sfGFP^{151,153}(AXC,AXC), or sfGFP^{149,153}(AXC,AXC) constructs, respectively. With dCNMO-dTPT3/NaMTP,TAT1TP, 81%, 89%, and 93% of the protein had two ncAAs with sfGFP^{149,151}(AXC,AXC), sfGFP^{151,153}(AXC,AXC), and sfGFP^{149,153}(AXC,AXC), respectively, while 19%, 11%, and 6% had a single ncAA. Cells transformed with the sfGFP^{149,151,153}(AXC,AXC,AXC) construct expressed protein that produced an even further shifted band, clearly indicating the incorporation of three AzK residues. Quantification of each band relative to total sfGFP revealed that the use of dNaM-dTPT3/NaMTP,TPT3TP resulted in 39%, 24%, and 33% of the protein having three, two and one ncAAs, respectively, and with fluorescence and protein shifts that were highly variable (Figure 7A and 7B). In contrast, use of the dCNMO-dTPT3/NaMTP,TAT1TP system resulted in 90% of the produced protein having all three ncAAs, with the remainder having two.

To further verify the successful incorporation of all three ncAAs with the dCNMO-dTPT3/NaMTP,TAT1TP system, we analyzed the isolated sfGFP by quantitative intact protein mass spectrometry. Briefly, purified proteins were desalted using centrifugal filter devices (Amicon® Ultra-0.5 – Millipore), and analyzed by HRMS (ESI-TOF). The mass spectra acquired were subsequently deconvoluted using the Waters MaxEnt 1 software, which proved to be quantitative upon peak integration (Figure S2). In agreement with the gel shift assay, this analysis revealed that that 88% of the isolated protein contained the expected three AzK residues, while the remaining 12% contained two AzK residues and a single Ile or Leu residue (Figure 7C).

DISCUSSION

The previously reported SSOs stored information with the UBPs dNaM-dTPT3, dPTMO-dTPT3, or dMTMO-dTPT3, and retrieved that information using NaMTP and TPT3TP. To explore the optimization of the SSO, we examined retention of the UBP, transcription into sfGFP mRNA and tRNA^{Pyl}, and decoding at the ribosome, using a collection of previously and newly reported deoxy- and ribonucleotide triphosphate analogs. We first examined the ability to store information with seven different dX-dTPT3 UBPs. In each case, the strand context of the UBP was the same, with dTPT3 and dX positioned in the corresponding antisense (template) strands of the sfGFP and tRNA^{Pyl} genes, respectively. With high concentrations of each dXTP provided, each dX-dTPT3 UBP is retained at a high level in the mRNA gene, with variation between 92% for dMTMOTP and 96% to 99% for dCNMOTP, dPTMOTP, and dNaMTP. Retentions in the tRNA gene were somewhat reduced, varying between 74% to 82%. As the concentrations of dXTP decreased, retentions remained roughly constant in the tRNA gene, but decreased in a dXTP specific manner in the mRNA gene, decreasing to 64% for dMTMOTP, but remaining relatively high, at ~94%,

for **dCNMOTP** and **d5FMTP**. The different concentration dependencies for retention in the tRNA and mRNA genes likely result from sequence context effects causing nucleotide insertion to be rate limiting in the mRNA and continued extension to be rate limiting in the tRNA gene. Exceptions were observed with **dNaMTP**, where at 10 μ M retention decreased to 73%, and as reported previously¹⁰ the cells did not survive when **dNaMTP** was provided at 5 μ M. In addition, with **dCNMOTP** and **d5FMTP**, retentions remained high in the mRNA (~93%) at even the lowest concentration examined. Loss of retention in the tRNA gene could result in less unnatural protein production, and perhaps more problematically, reduced fidelity of ncAA incorporation due to increased competition for decoding of the unnatural codon by “near-cognate” natural tRNAs. However, the fidelity of ncAA incorporation is clearly correlated with retention in the mRNA gene (Figure S3). Thus, the data demonstrates that each **dX** templates transcription of tRNA with sufficient efficiency and fidelity to not limit the fidelity of unnatural protein production. Based on these results, as well as those previously published²⁰, **d5FM-dTPT3**, and especially **dCNMO-dTPT3** are the most optimized UBPs, with their utility relative to **dNaM-dTPT3** deriving principally from their higher retention and protein production at lower unnatural triphosphate concentrations.

We next examined the ability of ten different **XTPs** to mediate the retrieval of information stored by the **dCNMO-dTPT3** UBP. Remarkably, all ten **XTPs** explored at high concentration are capable of mediating the production of proteins with at least moderate ncAA incorporation fidelity (98% to 63%). Generally, as **XTP** concentrations are decreased, the fidelity of ncAA incorporation decreased, indicating a reduction in the fidelity with which the unnatural mRNA is transcribed. This is consistent with the significant fluorescence observed when **XTP** was withheld.

The current study provides the first SARs for the transcription and translation of predominantly hydrophobic ribonucleotides. With the **XTPs**, when considering **NaMTP**, **PTMOTP**, and **MTMOTP**, it is clear that ring contraction and/or heteroatom derivatization is deleterious. However, with the monocyclic nucleobase **XTPs**, with the exception of **CIMOTP**, higher fidelity protein production is observed, relative to **MTMOTP** and **PTMOTP**, suggesting that the effects are more complicated than just aromatic surface area or heteroatom derivatization. It is likely that specific interactions between the unnatural nucleobases or with the polymerase are critical. Substitution at both the 4- and 5-positions of the monocyclic nucleobases has significant effects. Compared to **2OMeTP**, a Cl or Br substituent at the 4-position (**CIMOTP** and **BrMOTP**) is modestly deleterious, while a methyl group at the 4-position (**MMO2TP**), reduces overall fluorescence, but with a significant increase in protein fidelity. A nitrile substituent at the 4-position (**CNMOTP**) results in the production of the greatest amount of unnatural protein, and also modestly increases the fidelity with which the ncAA is incorporated. Addition of a fluoro substituent at the 5-position (**5F2OMeTP**) also increases both protein production and fidelity, relative to **2OMeTP**. The effects of substitution at the 4- and 5- positions appear at least approximately additive, as combining the 5-fluoro and 4-methyl substituents (**5FMTP**) allows for the high yield production of pure unnatural protein at lower concentrations, relative to the other unnatural triphosphates. However, while requiring higher concentrations, **NaMTP** provides the most optimal combination of yield and purity of the **XTP** analogs examined.

It is interesting to note that the SARs derived for the **XTP** analogs are distinctly different from those derived from the replication of **dXTP** analogs. For example, while **dPTMOTP**, **dCIMOTP**, and especially **dCNMO** are more optimal than **dNaMTP**, **CIMOTP** and **PTMOTP** are modestly and significantly less optimal than **NaMTP**. Moreover, while **dCNMO** is the most optimal partner for **dTPT3** discovered to date, the use of **CNMOTP** results in slightly reduced fidelity of ncAA incorporation, relative to **NaMTP**, although its use does result in the most unnatural protein production.

At the highest concentration, all five **YTPs** explored were effectively incorporated into the anticodon of tRNA^{PyI} and capable of mediating the production of proteins with at least moderate ncAA incorporation fidelity (98% to 51%). However, the UBP retention data suggests that the fidelity of protein production is not sensitive to modest loss of unnatural tRNA, implying that for the **YTPs** that resulted in lower protein gel shifts, transcription of the tRNA was either very inefficient or low fidelity. **TPT3TP**, **SICSTP**, **FSICSTP**, and **TAT1TP** were all, at least slightly toxic at the highest concentration (Table S2), and bulk cell fluorescence increased with decreasing concentrations. Unlike with **XTPs** and transcription of the mRNA, fidelity of unnatural protein did not decrease, suggesting that the increased protein production was simply the result of increased cell growth. In contrast, **5SICSTP** is not toxic (Table S2), and both unnatural protein production and fidelity decreased with decreasing concentrations, again presumably due to significantly compromised unnatural tRNA production.

When considering this **YTP** SAR, and using **SICSTP** as a reference, it is clear that a 7-fluoro substituent (**FSICSTP**) is quite deleterious, virtually ablating protein production, either due to effects on tRNA transcription or on translation, and only a small amount of protein is produced and with low fidelity. Addition of a methyl group at the 5-position (**5SICSTP**) reduces toxicity, but also appears to significantly reduce unnatural tRNA production. Ring contraction and heteroatom derivatization (**TPT3TP**) greatly improves both protein production and fidelity, but at high concentration it is the most toxic of the **YTP** analogs. Further heteroatom derivatization of the thiophene ring of **TPT3TP**, to produce the thiazole of **TAT1TP**, results in the production of even more pure unnatural protein than does **TPT3TP**, and importantly, with a substantial reduction in toxicity. Interestingly, unlike the case with the **XTP** SARs discussed above, the **YTP** SARs are relatively similar to those characterized with **dYTP** analogs.⁵ Given both the amount of protein produced and the fidelity of ncAA incorporation, **TAT1TP** is the most promising **YTP** identified to date.

Overall, the SARs identify the **dCNMO-dTPT3/NaMTP,TAT1TP** system as the most optimized for protein production, producing high amounts of protein with high fidelity incorporation of the ncAA. The use of **dCNMO-dTPT3/5FMTP,TAT1TP** produces protein with the same high fidelity, and while it produces slightly less protein, it requires the use of significantly less of the unnatural ribonucleotides. The utility of the **dCNMO-dTPT3/NaMTP,TAT1TP** system, relative to the previously reported **dNaM-dTPT3/NaMTP,TPT3TP** system, is particularly apparent with the encoding and retrieval of higher density unnatural information. As would be expected based on the results of the single labeled proteins, both systems produced protein with two ncAAs with high fidelity, but the

dCNMO-dTPT3/NaMTP,TAT1TP system generally produced the desired protein in greater quantities. Moreover, when encoding three ncAAs, the dNaM-dTPT3/NaMTP, TPT3TP system produced triply labeled protein with significantly reduced and more variable fidelities and yields, while the fidelity and yield with the dCNMO-dTPT3/NaMTP,TAT1TP system remained reproducibly high. The contaminant, where an Ile or Leu replaced a single ncAA, is unlikely to result from unnatural tRNA production, as UBP retention in the tRNA gene was high and similar for the both systems, and even with small differences, the single ncAA-incorporation data suggest that they should not cause significant reductions in the fidelity of unnatural protein production. It is also unlikely to result from mRNA transcription, which should be identical for the two systems (in both cases the dTPT3 directs the incorporation of NaM into the mRNA). Thus, the origin of the Leu/Ile contaminant is likely to be loss of the UBP in the mRNA gene during replication (which, as mentioned above, we were unable to directly measure). This is also consistent with the most common mutation expected, (dX to dT), which would produce an Ile codon.

CONCLUSION

This study reports the first SARs underlying efficient transcription by T7 RNAP and translation at the ribosome in the *E. coli* based SSO. It is particularly interesting that the X and dX SAR appear unrelated, while the Y and dY SAR appear more similar; it is possible the replicative polymerase(s), likely Pol III and to a lesser extent Pol II,¹¹ and T7 RNAP recognize distinct aspects of the (d)X nucleobases, but similar aspects of the (d)Y nucleobases. While the origins of this recognition will be pursued in future studies, the results have already identified a more optimal SSO, specifically an SSO that stores information with dCNMO-dTPT3, and retrieves the information it makes available with TAT1TP and NaMTP or 5FMTP. The optimization of the SSO is particularly apparent in its ability to produce protein with a higher density of ncAAs. The optimized SSO further attests to the ability of hydrophobic and packing forces to replace the complementary hydrogen bonds that underlie the storage and retrieval of natural information and represents significant progress towards the creation of a SSO with a fully unrestricted expansion of its genetic alphabet and code.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

This work was supported by the National Institutes of Health (GM118178 to F.E.R. and GM128376 to R.J.K.). A.W.F. was supported by a National Science Foundation Graduate Research Fellowship (NSF/DGE-1346837). E.C.F. was supported by a Boehringer Ingelheim Fonds PhD Fellowship. B.A.A. was supported by NASA Exobiology (NNX14AP59G to R.K.). L.L. was supported by cooperation findings from Henan Normal University (NSFC21472036).

REFERENCES

1. Leduc S, The Mechanisms of Life Rebman Company: New York, 1911.

2. Yang Z; Chen F; Alvarado JB; Benner SA, Amplification, Mutation, and Sequencing of a Six-letter Synthetic Genetic System. *J. Am. Chem. Soc* 2011, 133, 15105–15112. [PubMed: 21842904]
3. Dhami K; Malyshev DA; Ordoukhanian P; Kubelka T; Hocek M; Romesberg FE, Systematic Exploration of a Class of Hydrophobic Unnatural Base Pairs Yields Multiple New Candidates for the Expansion of the Genetic Alphabet. *Nucleic Acids Res* 2014, 42, 10235–10244. [PubMed: 25122747]
4. Hirao I; Kimoto M, Unnatural Base Pair Systems Toward the Expansion of the Genetic Alphabet in the Central Dogma. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci* 2012, 88, 345–367.
5. Feldman AW; Romesberg FE, Expansion of the Genetic Alphabet: A Chemist's Approach to Synthetic Biology. *Acc. Chem. Res* 2018, 51, 394–403. [PubMed: 29198111]
6. Ast M; Gruber A; Schmitz-Esser S; Neuhaus HE; Kroth PG; Horn M; Haferkamp I, Diatom Plastids Depend on Nucleotide Import from the Cytosol. *Proc. Natl. Acad. Sci. USA* 2009, 106, 3621–3626. [PubMed: 19221027]
7. Malyshev DA; Dhami K; Lavergne T; Chen T; Dai N; Foster JM; Correa IR Jr.; Romesberg FE, A Semi-synthetic Organism with an Expanded Genetic Alphabet. *Nature* 2014, 509, 385–388. [PubMed: 24805238]
8. Zhang Y; Lamb BM; Feldman AW; Zhou AX; Lavergne T; Li L; Romesberg FE, A Semisynthetic Organism Engineered for the Stable Expansion of the Genetic Alphabet. *Proc. Natl. Acad. Sci. USA* 2017, 114, 1317–1322. [PubMed: 28115716]
9. Zhang Y; Ptacin JL; Fischer EC; Aerni HR; Caffaro CE; San Jose K; Feldman AW; Turner CR; Romesberg FE, A Semi-synthetic Organism that Stores and Retrieves Increased Genetic Information. *Nature* 2017, 551, 644–647. [PubMed: 29189780]
10. Dien VT; Holcomb M; Feldman AW; Fischer EC; Dwyer TJ; Romesberg FE, Progress Toward a Semi-Synthetic Organism with an Unrestricted Expanded Genetic Alphabet. *J. Am. Chem. Soc* 2018, 140, 16115–16123. [PubMed: 30418780]
11. Ledbetter MP; Karadeema RJ; Romesberg FE, Reprogramming the Replisome of a Semi-Synthetic Organism for the Expansion of the Genetic Alphabet. *J. Am. Chem. Soc* 2018, 140, 758–765. [PubMed: 29309130]
12. Wang L; Brock A; Herberich B; Schultz PG, Expanding the Genetic Code of *Escherichia coli*. *Science* 2001, 292, 498–500. [PubMed: 11313494]
13. Wang Q; Xie XY; Han J; Cui G, QM and QM/MM Studies on Excited-State Relaxation Mechanisms of Unnatural Bases in Vacuo and Base Pairs in DNA. *J. Phys. Chem. B* 2017, 121, 10467–10478. [PubMed: 29083167]
14. Negi I; Kathuria P; Sharma P; Wetmore SD, How do Hydrophobic Nucleobases differ from natural DNA Nucleobases? Comparison of Structural Features and Duplex Properties from QM Calculations and MD Simulations. *Phys. Chem. Chem. Phys* 2017, 19, 16365–16374. [PubMed: 28657627]
15. Jahiruddin S; Mandal N; Datta A, Structure and Electronic Properties of Unnatural Base Pairs: The Role of Dispersion Interactions. *Chemphyschem* 2018, 19, 67–74. [PubMed: 29139595]
16. Jahiruddin S; Datta A, What Sustains the Unnatural Base Pairs (UBPs) with no Hydrogen Bonds. *J Phys Chem B* 2015, 119, 5839–5845. [PubMed: 25893481]
17. Guo WW; Zhang TS; Fang WH; Cui G, QM/MM studies on the Excited-state Relaxation Mechanism of a Semisynthetic dTPT3 base. *Phys Chem Chem Phys* 2018, 20, 5067–5073. [PubMed: 29388994]
18. Galindo-Murillo R; Barroso-Flores J, Structural and Dynamical Instability of DNA caused by high occurrence of d5SICS and dNaM unnatural nucleotides. *Phys. Chem. Chem. Phys* 2017, 19, 10571–10580. [PubMed: 28394373]
19. Bhattacharyya K; Datta A, Visible-Light-Mediated Excited State Relaxation in Semi-Synthetic Genetic Alphabet: d5SICS and dNaM. *Chemistry* 2017, 23, 11494–11498. [PubMed: 28675549]
20. Feldman AW; Romesberg FE, In Vivo Structure-Activity Relationships and Optimization of an Unnatural Base Pair for Replication in a Semi-Synthetic Organism. *J. Am. Chem. Soc* 2017, 139, 11427–11433. [PubMed: 28796508]
21. Pedelacq JD; Cabantous S; Tran T; Terwilliger TC; Waldo GS, Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol* 2006, 24, 79–88. [PubMed: 16369541]

22. Srinivasan G; James CM; Krzycki JA, Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science* 2002, 296, 1459–1462. [PubMed: 12029131]
23. Polycarpo CR; Herring S; Berube A; Wood JL; Soll D; Ambrogelly A, Pyrrolysine Analogues as Substrates for Pyrrolysyl-tRNA Synthetase. *FEBS Lett* 2006, 580, 6695–6700. [PubMed: 17126325]
24. Nguyen DP; Lusic H; Neumann H; Kapadnis PB; Deiters A; Chin JW, Genetic encoding and labeling of aliphatic azides and alkynes in recombinant proteins via a pyrrolysyl-tRNA Synthetase/tRNA(CUA) pair and click chemistry. *J. Am. Chem. Soc* 2009, 131, 8720–8721. [PubMed: 19514718]
25. Chatterjee A; Sun SB; Furman JL; Xiao H; Schultz PG, A Versatile Platform for Single- and Multiple-Unnatural Amino acid Mutagenesis in *Escherichia coli*. *Biochemistry* 2013, 52, 1828–1837. [PubMed: 23379331]
26. Seo YJ; Malyshev DA; Laverne T; Ordoukhanian P; Romesberg FE, Site-specific labeling of DNA and RNA using an efficiently replicated and transcribed class of unnatural base pairs. *J. Am. Chem. Soc* 2011, 133, 19878–19888. [PubMed: 21981600]
27. Baskin JM; Prescher JA; Laughlin ST; Agard NJ; Chang PV; Miller IA; Lo A; Codelli JA; Bertozzi CR, Copper-free Click Chemistry for Dynamic in vivo Imaging. *Proc. Natl. Acad. Sci. USA* 2007, 104, 16793–16797. [PubMed: 17942682]
28. Ludwig J; Eckstein F, Rapid and Efficient Synthesis of Nucleoside 5'-O-(1-Thiotriphosphates), 5'-Triphosphates and 2',3'-Cyclophosphorothioates using 2-Chloro-4H-1,3,2-benzodioxaphosphorin-4-one. *J. Org. Chem* 1989, 54, 631–635.
29. Seo YJ; Matsuda S; Romesberg FE, Transcription of an Expanded Genetic Alphabet. *J. Am. Chem. Soc* 2009, 131, 5046–5047. [PubMed: 19351201]

Figure 1.
The dNaM-d5SICS, dNaM-dTPT3, dCNMO-dTPT3, and dPTMO-dTPT3 UBPs.

Figure 2.
dXTP analogs. Ribose and phosphates omitted for clarity.

Figure 3.

Optimization of translation to incorporate AzK into sfGFP using various dXTPs. **(A)** UBP retention (%) in the sfGFP gene. **(B)** UBP retention (%) in the tRNA^{Pyl} gene. **(C)** Relative sfGFP fluorescence observed normalized to cell growth (Relative Fluorescence Units (RFU) per OD₆₀₀) in the presence or absence of AzK. **(D)** Protein shift (%) measured by western blot. Each bar represents the mean, with error bars indicating standard error ($n = 3$). Open circles represent data for each independent trial. Asterisk indicates that cells were unable to grow under the condition indicated.

Figure 4. Ribonucleotide analogs. (A) XTP analogs. (B) YTP analogs. Ribose and phosphates omitted for clarity.

Figure 5.

SAR analysis of translation using various unnatural ribonucleotides to incorporate AzK into sfGFP. (A) Total sfGFP fluorescence (RFU) observed in the presence of AzK for **XTP** analogs, where (-) represents control samples provided with only **TPT3TP** (**XTP** withheld). (B) Protein shift (%) measured by western blot for **XTP** analogs, where (-) represents control samples provided with only **TPT3TP** (**XTP** withheld). (C) Total sfGFP fluorescence (RFU) observed in the presence of AzK for **YTP** analogs, where (-) represents control samples provided with only **NaMTP** (**YTP** withheld). (D) Protein shift (%) measured by western blot for **YTP** analogs, where (-) represents control samples provided with only **NaMTP** (**YTP** withheld). Each bar represents the mean, with error bars indicating standard error (n = 4). Open circles represent data for each independent trial.

Figure 6. Optimization of unnatural ribonucleotide triphosphate concentrations. (A) Total sfGFP fluorescence (RFU) as a function of the concentrations of NaMTP and TAT1TP (μM). (B) Total sfGFP fluorescence (RFU) as a function of the concentrations of 5FMTP and TAT1TP (μM). (C) Protein shift (%) as a function of the concentrations of NaMTP and TAT1TP (μM). (D) Protein shift (%) as a function of the concentrations of 5FMTP and TAT1TP (μM). Error bars indicate standard error of each value ($n = 3$).

Figure 7.

Storage and retrieval of higher density unnatural information with either dNaM-dTPT3/NaMTP,TPT3TP or dCNMO-dTPT3/NaMTP,TAT1TP. (A) Total sfGFP fluorescence (RFU) observed in the presence of AzK. (B) Protein shift (%) measured by western blot. For strip charts, each bar represents the mean, with error bars indicating standard error ($n = 4$), and open circles represent data for each independent trial. (C) Representative spectrum of quantitative HRMS analysis of triple labeled protein produced using the dCNMO-dTPT3/NaMTP,TAT1TP. Peak labels show deconvoluted molecular weight of intact protein, with amino acid residues at positions 149, 151, and 153 shown and quantification of each peak (% , $n = 3$) shown below. See Supporting Information for assignment of unlabeled peaks.