Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Expansion of the genetic code via expansion of the genetic alphabet

Vivian T Dien[1], Sydney E Morris[1], Rebekah J Karadeema[1] and Floyd E Romesberg

Current methods to expand the genetic code enable site-specific incorporation of non-canonical amino acids (ncAAs) into proteins in eukaryotic and prokaryotic cells. However, current methods are limited by the number of codons possible, their orthogonality, and possibly their effects on protein synthesis and folding. An alternative approach relies on unnatural base pairs to create a virtually unlimited number of genuinely new codons that are efficiently translated and highly orthogonal because they direct ncAA incorporation using forces other than the complementary hydrogen bonds employed by their natural counterparts. This review outlines progress and achievements made towards developing a functional unnatural base pair and its use to generate semi-synthetic organisms with an expanded genetic alphabet that serves as the basis of an expanded genetic code.

Address
Department of Chemistry, The Scripps Research Institute, La Jolla, CA 92037, USA

Corresponding author: Romesberg, Floyd E (floyd@scripps.edu)
[1] These authors contributed equally.

Current Opinion in Chemical Biology 2018, **46**:196–202

This review comes from a themed issue on **Synthetic biology**

Edited by **Michael P Ledbetter** and **Floyd E Romesberg**

https://doi.org/10.1016/j.cbpa.2018.08.009

Introduction

Biological diversity allows life to adapt to different environments, and over time, evolve new forms and functions. The source of this diversity is the variation within protein sequences provided by the twenty natural amino acids, variation that is encoded in an organism's genome by the four natural DNA nucleotides. Although the functional diversity provided by the natural amino acids may be high, the vastness of sequence space dramatically limits what might actually be explored, and moreover, some functionality is simply not available. Nature's use of cofactors for hydride transfer, redox activity, electrophilic bond formation and so on, attests to these limitations. Furthermore, with the increasing focus on developing proteins as therapeutics [1], these limitations are problematic, as the physiochemical diversity of the natural amino acids is dramatically restricted compared to that of the small molecule drugs designed by chemists. In principle, it should be possible to circumvent these limitations by expanding the genetic code to include additional, non-canonical amino acids (ncAAs) with desired physiochemical properties.

Almost 20 years ago, Peter Schultz increased the diversity available to living organisms by expanding the genetic code using the amber stop codon (UAG) to encode ncAAs in *Escherichia coli* [2••,3••]. This landmark accomplishment was achieved using a tRNA–amino acid tRNA synthetase (aaRS) pair from *Methanococcus jannaschii*, in which the tRNA was recoded to suppress the stop codon and the aaRS was evolved to charge the tRNA with an ncAA. This method of codon suppression has since been expanded to the other stop codons [4] and even quadruplet codons [5], as well as to the use of several other orthogonal tRNA–aaRS pairs (most notably the pyrrolysyl (Pyl) tRNA–synthetase pair from *Methanosarcina barkeri/mazei* [6,7,8•,9]), broadening the scope of ncAAs that may be incorporated into proteins. These methods have already begun to revolutionize both chemical biology [10–12] and protein therapeutics [13].

Though these methods enable incorporation of up to two, different ncAAs in both prokaryotic [14] and eukaryotic [15] cells, the heterologous recoded tRNAs must compete with endogenous release factors (RFs), or in the case of quadruplet codons, normal decoding [16], which limits the efficiency and fidelity of ncAA incorporation. To eliminate competition with RF1, which recognizes the amber stop codon and terminates translation, efforts have been directed toward removal of many or all instances of the amber stop codon in the host genome [17,18] or modification of RF2 [19] to allow for the deletion of RF1. However, eukaryotes have only one release factor, and while it may be modified [20] it cannot be deleted, and with prokaryotes, deletion of RF1 results in greater mis-suppression of the amber stop codon by other tRNAs, which reduces the fidelity of ncAA incorporation [21]. Though Herculean efforts to further exploit codon redundancy to liberate natural codons for reassignment to ncAAs are underway [22], codon reassignment may be complicated by pleiotropic effects, as codons are not truly redundant, for example due to their effects on the rate of translation and protein folding [23]. In addition, codon

reappropriation is limited by the challenges of large-scale genome engineering, especially for eukaryotes [24].

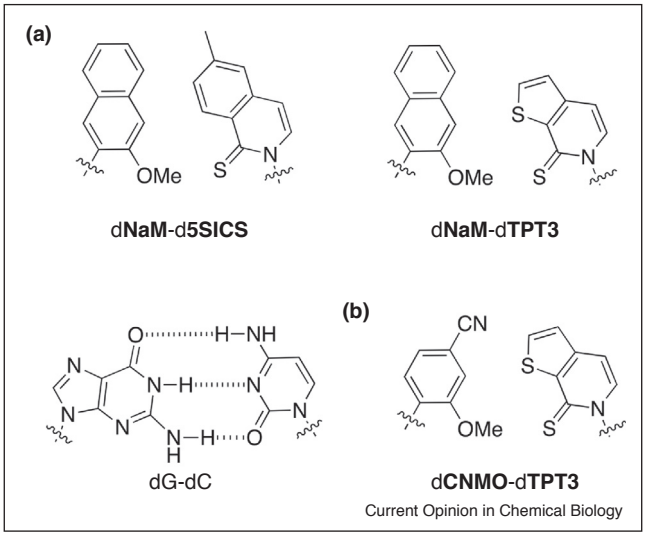
An alternative approach to natural codon reassignment is the creation of entirely new codons that are free of any natural function or constraint, and whose recognition at the ribosome is inherently more orthogonal. This may be accomplished through the creation of organisms that harbor a fifth and sixth nucleotide that form an unnatural base pair (UBP). Such semi-synthetic organisms (SSOs) would need to faithfully replicate DNA containing the UBP, efficiently transcribe it into mRNA and tRNA containing the unnatural nucleotides, and then efficiently decode unnatural codons with cognate unnatural anticodons. Such SSOs would have a virtually unlimited number of new codons to encode ncAAs.

Development of a functional UBP

The first challenge in developing an expanded genetic alphabet is the identification of unnatural nucleotides that selectively pair in duplex DNA and during replication by a DNA polymerase. Although the Benner group (who has contributed to this special issue) has approached this challenge with synthetic nucleotides that pair via hydrogen bonding (H-bonding) patterns that are orthogonal to the natural base pairs [25], our group, as well as the Hirao group (who has also contributed to this special issue), took inspiration from the demonstration by the Kool group that hydrogen bonds are not required for DNA polymerase-mediated insertion of a triphosphate [26,27]. Although the Hirao group has pursued shape analogs of the natural nucleobases [28], we have pursued the *de novo* development of a UBP, starting from simple aromatic nucleobase analogs with the ability to pair via hydrophobic and packing forces. We have taken a medicinal chemistry-like approach to guide the exploration of nearly 200 analogs, with design and analysis proceeding iteratively via the generation of structure–activity relationships (SARs).

These medicinal chemistry-like efforts have been reviewed [29•] and they ultimately culminated in the discovery of a family of UBPs that were well retained during PCR amplification [30,31]. Several key and interesting SARs are of note, but perhaps the most important is related to the nucleobase substituent *ortho* to the glycosidic bond. The data clearly demonstrate that efficient unnatural triphosphate insertion and continued extension of the growing strand of DNA require an *ortho* substituent that is capable of both hydrophobic packing and accepting an H-bond. This apparent physiochemical contradiction appears to have been resolved by the thioamide and methoxy substituents, as found in the d**NaM**-d**5SICS** and d**NaM**-d**TPT3** UBPs (Figure 1a). The polarizability of the sulfur increases its ability to pack, relative to an oxygen, without ablating its ability to accept an H-bond, and the methoxy substituent, via simple bond rotation,

Figure 1



Base pairs. **(a)** The d**NaM**-d**5SICS** and d**NaM**-d**TPT3** UBPs and a natural dG-dC pair (with complementary H-bonding shown). **(b)** The d**CNMO**-d**TPT3** UBP. Phosphate and sugar moieties omitted for clarity.

can present either a hydrophobic methyl group or a more hydrophilic ether oxygen.

In addition to the nature of the *ortho* substituent, SAR data revealed that nucleobase aromatic surface area extending into the developing major groove favors both triphosphate insertion and extension, presumably through optimized packing interactions (Figure 2). However, careful optimization was required to favor packing between the unnatural nucleobase and the nucleobase of the primer terminus, as opposed to packing with the unnatural and natural nucleobases in the template strand via cross-strand intercalation [32,33]. In free duplex DNA, the UBPs adopt a cross-strand intercalated structure, but remarkably, in the polymerase active site when the unnatural triphosphate is paired opposite its cognate unnatural nucleotide in the template, a Watson–Crick-like structure is adopted [34,35]. Thus, DNA containing the UBP is replicated with a mutually induced-fit mechanism, wherein formation of the UBP drives the same conformational change required to form the closed polymerase complex that is induced by the formation of a natural base pair, and the tightly packed environment of the closed polymerase drives the UBP to adopt the required (de-intercalated) structure. However, after the polymerase translocates to position the next templating nucleotide in the active site, the nascent UBP again cross-stand intercalates, which distorts the primer terminus and mandates de-intercalation for efficient continued primer elongation. Using this structural data as a guide, we contracted and derivatized the d**5SICS** nucleobase with a sulfur atom to yield d**TPT3**, which appears to favor de-intercalation and intrastrand packing and which