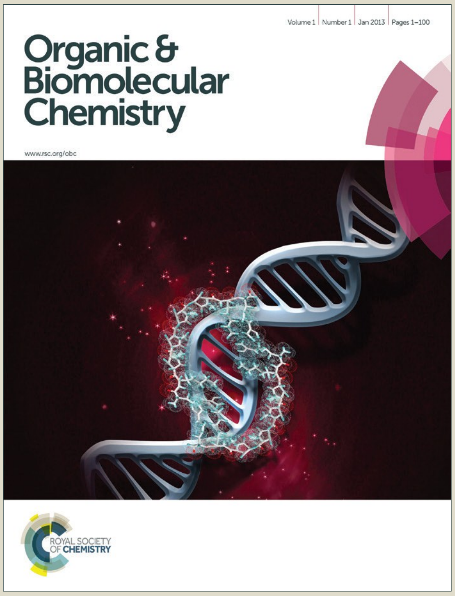
Organic & Biomolecular Chemistry

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Organic & Biomolecular Chemistry
www.rsc.org/obc
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/obc

Journal Name RSCPublishing

ARTICLE

Discovering functional, non-proteinogenic amino acid containing, peptides using genetic code reprogramming

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

J. M. Rogers and H. Suga*

The protein synthesis machinery of the cell, the ribosome and associated factors, is able to accurately follow the canonical genetic code, that which maps RNA sequence to protein sequence, to synthesize functional proteins from the twenty or so proteinogenic amino acids. A number of innovative methods have arisen to take advantage of this accurate, and efficient, machinery to direct the assembly of non-proteinogenic amino acids. We review and compare these routes to 'reprogram the genetic code' including *in vitro* translation, engineered aminoacyl tRNA synthetases, and RNA 'flexizymes'. These studies show that the ribosome is highly tolerant of unnatural amino acids, with hundreds of unusual substrates of varying structure and chemistries being incorporated into protein chains. We also discuss how these methods have been coupled to selection techniques, such as phage display and mRNA display, opening up an exciting new avenue for the production of proteins and peptides with properties and functions beyond that which is possible using proteins composed entirely of the proteinogenic amino acids.

Introduction

Life has evolved highly sophisticated machinery to reproducibly synthesize linear, chemically heterogeneous polymers, such as RNA, DNA and proteins. The cellular function of these polymers relies on the appropriate monomers being assembled in a prescribed sequence. The information to direct this assembly is ultimately derived from the genetic material of the organism - itself a polymer (DNA). Using the sequence information of one polymer to direct the assembly of another relies on robust genetic codes. This code is straightforward for the assembly of RNA and DNA: the helical structures of nucleic acid polymers[1] and the energetics of base pairing guide the assembly of complementary DNA and RNA strands. The case is very different for the synthesis of proteins and peptides, which is guided by the sequence of a 'messenger' mRNA. There are a larger number of monomers, 20+ amino acids vs. 4 nucleic acid bases, and the dissimilar chemistry of the amino acid monomers and the RNA, prevents a direct one-to-one interaction, like base pairing, to guide this assembly. Protein synthesis requires a complex code[2] and highly sophisticated machinery (40+ components)[3]. The cellular machinery, the ribosome and associated factors, that carries out protein assembly, 'translation', is highly conserved in evolution[4]. Despite this complexity, translation can be achieved *in vitro*, *i.e.* in the test tube, using purified components or cell extracts[3].

Recent work has shown that this translation machinery is surprisingly amenable to manipulation. Here we describe recent attempts to alter this machinery *in vitro* and *in vivo* to 'reprogram' the genetic code. This reprogramming allows for the incorporation of non-proteinogenic amino acids (npAA), alongside the proteinogenic amino acids (pAA). We describe how these powerful techniques can be applied to discover artificial proteins and peptides with novel functions.

Why reprogram the genetic code?

For many applications it is possible, and advantageous, to rationally design and synthesize npAA containing peptides. Examples of current importance include hydrocarbon stapling[5] and foldamer-peptide hybrids[6]. If npAA containing peptides can be successfully synthesized *in vitro*, why go to the effort of reprogramming the genetic code? In the particular case of *in vivo* genetic code reprogramming, this allows for the study of npAA containing proteins with controllable chemical reactivity, specific post-translational modifications, and useful optical and magnetic probes in the context of the living cell. These applications are extensively reviewed elsewhere[7]. The other main motivation for genetic code reprogramming, which will be the focus of this review, is that it allows the incorporation of npAA during 'selections' for peptides and proteins with novel functions.

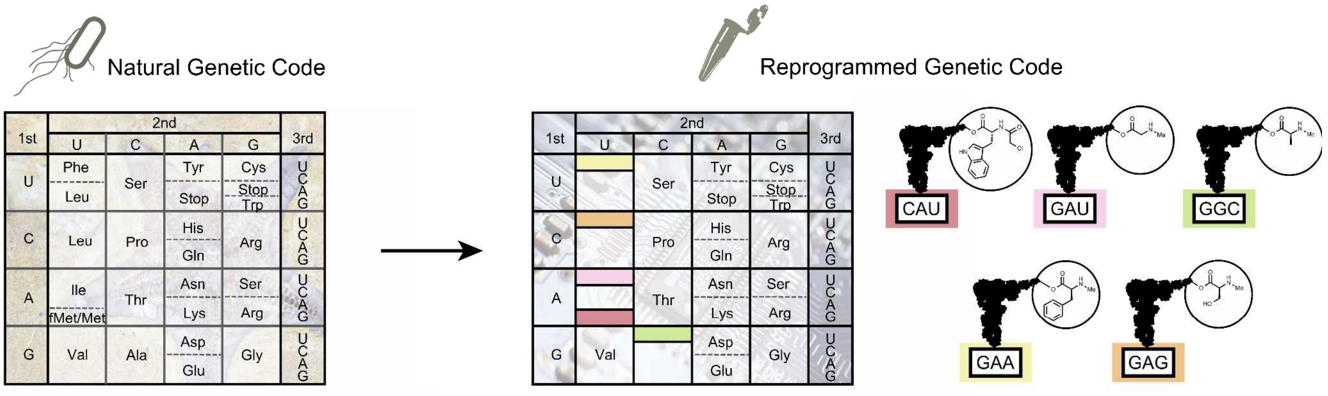


Fig. 1 Genetic code reprogramming. The natural genetic code, essentially the same in all organisms, maps 'codons' (three letter strings, where each letter is one of the four base pairs mRNA) onto one of the 20 proteinogenic amino acid (pAAs). While the code is redundant, i.e. many codons code for the same amino acid, all codons are in use. The first step in genetic code reprogramming is to vacate a codon; the second step is to add a tRNA with the corresponding anticodon, loaded with the chosen non-proteinogenic amino acid (npAA). Shown is the reprogrammed genetic code used by Yamagishi *et al.*[8]: the codon for methionine is reprogrammed to the npAA 2-chloroacetyl (ClAc) D-tryptophan, phenylalanine to N-methyl phenylalanine, leucine to N-methyl serine, isoleucine to N-methyl serine, and alanine to N-methyl alanine. Left photo by Didier Descouens, Creative Commons Attribution-Share Alike 3.0 Unported licence, right photo by Harland Quarrington/MOD, Open Government Licence v1.0.

The first step of selection is to create a large library of randomized DNA sequences. These are commercially available and can be synthesized quickly and cheaply. Such libraries can be used in combination with genetic code reprogramming to produce enormous numbers (billions to trillions) of unique, npAA containing peptides. If these peptides are linked (see below) to the RNA/DNA coding for their sequence, rounds of competition for a particular function can 'select' a functional peptide(s). Most frequently this function is to bind to a chosen biological target - the library of peptides can be incubated with immobilized target and the weaker binding peptides can be washed away. The precise composition of the tighter binding peptides can be confirmed after recovery by sequencing the corresponding nucleic acid region which, using knowledge of the altered genetic code, can be converted into peptide primary structure. The discovered peptides may then affect the function of the target, *e.g.* they may act as useful inhibitors.

Benefits of non-proteinogenic amino acids

The properties of the peptide backbone and the pAA's side-chains engender natural proteins with the ability to fold to a specific three-dimensional structure, interact specifically with other biological macromolecules, perform mechanical roles or act as enzymes. Peptides suitable for solid phase chemical synthesis are often shorter than these natural proteins (consisting of tens, rather than hundreds, of amino acids) and, as a result, are likely to be disordered in isolation, occupying a large number of rapidly interconverting conformations[9]. This is not an obstacle to binding, and many naturally occurring proteins undergo a disorder-order transition upon binding[10], with some degree of folding can occur upon interaction[11]. While this allows for relatively small lengths of peptide chain to interact with a large surface area[12], binding is generally weaker than the interactions between folded proteins[13].

Tight binding of short peptides to a specific target can often be achieved through the incorporation of npAAs. While specific preformed structure is not necessarily required for binding[14], npAA that can reduce the accessible conformations in the unbound state can reduce the entropic penalty inherent in these disorder-order transitions. One highly effective method to achieve this is to convert the peptide into a macrocycle, cyclizing part, or all, of the otherwise linear chain. Incorporating npAA with appropriate reactivity can allow for efficient, spontaneous macrocyclization using a non-reducible covalent bond[15].

If the peptide is to be used in a biological setting (in a serum, with model organisms or as a drug) using only pAAs may result in a molecule that is highly susceptible to degradation by endogenous proteases[16], be unable to permeate the cellular membrane to reach the desired target, or, for those intended as drugs, may not be orally available. If npAA permit macrocyclization, this can prevent degradation by proteases, as cyclization can disfavour the extended conformations required for recognition by these enzymes[17]. Introduction of certain npAAs is another way to make protease recognition more difficult, particularly if there is a modification close to the backbone, such as N-methylation. N-methylation has received much attention as it has the added benefit of reducing the number of hydrogen bond donors, often with the consequence of aiding diffusion across the cellular membrane and enhancing oral availability[18]. The combination of macrocyclization and N-methylation is also used by naturally occurring, functional peptides such as the widely used immunosuppressant cyclosporin[19].

Genetic code reprogramming

The fidelity of the natural genetic code relies on correct aminoacylation of transfer RNAs (tRNAs) by amino acids, catalysed by aminoacyl tRNA synthetases (aaRSs), and the specific antiparallel codon–anticodon recognition between the mRNA inside the ribosome and these tRNAs. This gives rise to the codon-amino acid table show in Fig. 1. All genetic code

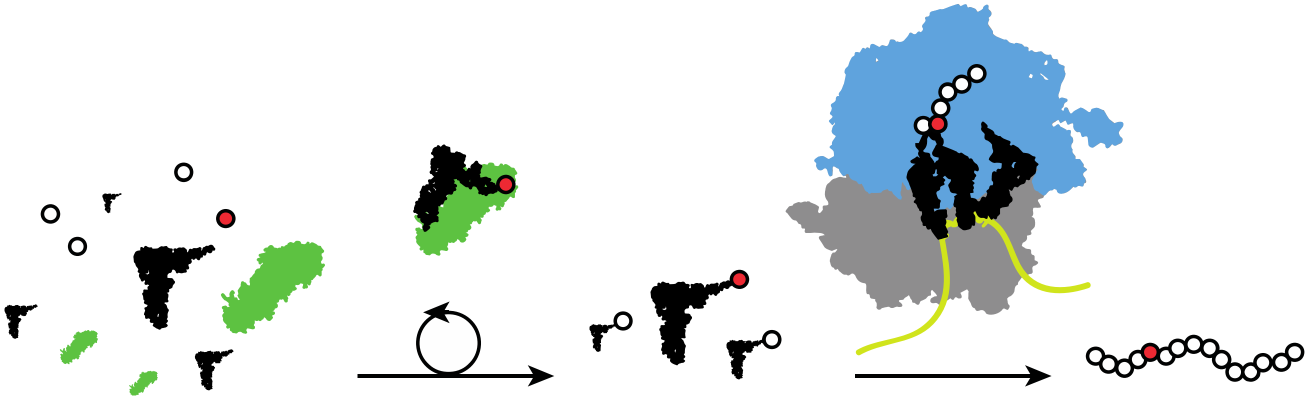


Fig. 2 npAA incorporation using aaRSs and *in vitro* translation. The first step of this simplified scheme shows the pAAs (white circle) and npAAs (red circles) being attached to tRNAs (black) by aaRSs (green). Each aaRS is specific for a number of tRNAs all of which are loaded with the same amino acid. Some npAAs are fortuitously substrates for the natural aaRSs, other npAA require engineered aaRSs. The second step of this scheme shows how, in the same reaction vessel/cell, these amino acylated tRNAs are used by the ribosome (blue and grey), directed by the mRNA (yellow), to form the final peptide (far right).

reprogramming methods have two features in common, first, codons must be vacated, or created, to avoid competition with the natural amino acid(s), second, npAAs must be attached to tRNAs with the corresponding anticodons (Fig. 1).

Early successes involved chemically attaching npAA to fragments of tRNA, followed by enzymatic ligation to make full length tRNA[20]. If this tRNA has an anticodon corresponding to the ‘empty’ codon UAG that would otherwise code for ‘stop’, loaded tRNA can compete with the natural release factor for incorporation. Alternatively, *in vitro,* release factors can be withdrawn. This labour intensive approach remains in use[21] but has been largely superseded by the following methods.

Aminoacylation using the intrinsic promiscuity of aaRS

The most direct method of genetic code reprogramming is to take advantage of the intrinsic promiscuity of the existing aaRSs which can, to some extent, aminoacylate tRNA using npAA. Even though the promiscuity of aaRSs is generally modest, an excess amount of npAA, ideally structurally similar to the pAA, can be tolerated for charging onto cognate tRNA *in vivo* as well as *in vitro*. Examples *in vivo* include adding selenomethionine to *E. coli* auxotrophic for methionine[22], to aid the structural elucidation of proteins by X-ray crystallography; and replacing hydrophobic amino acids with fluorinated homologues[23], to produce proteins with altered biophysical properties.

In vitro, a greater degree of reprogramming, replacing a larger number of pAAs with more chemically diverse npAA can be achieved. The development of the PURE *in vitro* translation system, using purified ribosomes, separately purified recombinant proteins (aaRSs, initiation, elongation and release factors), and required small molecules (such as free amino acids, ATP), has greatly facilitated such efforts[3]. Codons can be vacated simply by not adding a subset of the pAAs to the translation mix. The Szostak laboratory have tested a large number of npAA and have shown that many, albeit those which in part structurally resemble the cognate pAAs, are suitable substrates for the wild-type aaRSs[24], When the cognate pAAs are withdrawn and replaced with such npAAs in the PURE system, incorporation of these npAAs into the peptide chain can be realized[25] (Fig. 2). Using this method they were able to reassign 13 codons to different npAAs, and show that these can be incorporated into a single peptide chain.

Novel, functional npAA-containing peptides have been discovered by combining this method of genetic code reprogramming with selections via ‘mRNA display’. mRNA display uses the properties of the antibiotic puromycin to form a covalent linkage between the mRNA and the newly synthesized peptide it encodes[26]. Using this method the Szostak lab tested 10^{13} peptides in parallel for binding to the protease thrombin. They discovered tight, (1.5 nM K_d) binders for this target and, importantly, the binding affinity was dependent on the encoded npAAs[27].

This method of genetic code reprogramming has been successful at specifically encoding a large number of chemically diverse npAAs. Of particular note is the incorporation of selenium containing npAA, as these can be converted to dehydroalanine by oxidative elimination, and then attacked by thiol containing amino acids such as cysteine. This allows for the formation of lanthionine-like thioether (sulfide) bridged macrocyclic peptides. This was used in a selection against the protein Sortase A to produce a macrocyclic peptide with modest (3 μM K_d) binding affinity[28].

The disadvantage of this method is that it relies on fortuitous, specific recognition between the npAA and an aaRS. If one npAA in particular is required, there is no guarantee that a suitable aaRS exist, and if it does, there is little choice over which codon to assign it to, and therefore the corresponding pAA that must be withdrawn.

Engineered aaRSs

Rather than relying on the promiscuity of the natural aaRSs, it is also possible to alter existing aaRSs to allow for the specific recognition of npAA substrates. Rational protein engineering and/or traditional (i.e. without npAAs) selection techniques can be used to discover mutant aaRS, and specialized tRNA that can

accept npAAs (often artificial, unnatural npAA) while not reacting with pAAs and wild-type aaRSs. These ‘orthogonal’ aaRS/tRNA pairs can be discovered by starting the engineering process with aaRS/tRNA from an organism distantly related to that which provides the other aaRSs. The use of engineered aaRS has been highly successful *in vivo* for a number of model organisms[29], and there have been some applications using *in vitro* translation systems[30]. When used *in vivo*, codons cannot be easily vacated, npAA are often assigned to the UAG stop codon or must be assigned a four-base codon and/or orthogonal ribosomes introduced[15c].

The fact that these engineered aaRS/tRNA pairs are suitable for use in *E. coli* allows for novel peptides containing npAAs to be discovered using phage display. In this technique, each potentially functional peptide is expressed on the surface of a separate virus particle, a particle with the DNA encoding for this peptide enclosed. As these two molecules are coupled, those tighter binding peptides can be isolated and the encoding DNA recovered[31]. The Schultz laboratory have used engineered aaRS to include npAA that provide antibodies with improved protein binding[32], and sugar binding[33]. More recently, they have included bidentate ligands for metal binding in a small peptide[34], and evolved a zinc-finger-like DNA binding protein than, instead of zinc, binds iron (II) in its structural core[35].

Another *in vivo* selection technique is to express the library of peptides in individual cells and couple the function of the peptide to the survival of these cells. Young *et al.* found peptide inhibitors of HIV-1 protease from a random pool, where each peptide has the potential to contain a ketone containing npAA and to be head-to-tail, backbone cyclized by a protein-catalysed split-intein system[36]. Each cell expressed one of the library peptides, HIV-1 protease and an antibiotic resistance gene containing a HIV-1 cleavage site: only if the macrocyclic, npAA-containing peptide inhibited HIV-1 protease could that cell survive. However, either due to the small size of the peptides, or the smaller cell-based library size, their IC_{50} values were only in the low μM range.

Frost *et al.* used engineered aaRSs to include a unnatural thiol containing npAA into peptides which, when coupled to the split-intein system, can generate alternative topology, i.e. not head-to-tail, macrocyclic peptides. With a 96-well plate format, ELISA-like selection scheme, they improved the binding of an existing streptavidin binding motif[37].

The main advantage of engineered aaRS mediated genetic code reprogramming is that particular npAAs and codon assignment can be chosen, and does not rely on fortuitous wild-type aaRS recognition. Using engineered aaRSs also means that all of the pAAs can still be used, which is important for making modifications to an existing peptide/protein. However, inspection of the published substrates suggests that there is some requirement, or at least preference, for structural similarity between the (to be engineered) aaRS’s cognate substrate and the chosen npAA[7b], presumably to increase the chances of obtaining an engineered aaRS with specific catalytic activity. Another advantage of working *in vivo* is that large folded proteins, such as antibody fragments, can be produced and in the cellular context where there are chaperones to aid folding. Additionally, phage display, as a classical selection technique, appears to be accessible for many laboratories.

Phage display, and other *in vivo* selection techniques, have the disadvantage of small library size (billions) compared to *in vitro* techniques such as mRNA display with orders of magnitude larger library size (trillions). When discovering peptides with completely novel functions, this smaller library size could be a significant disadvantage.

In vivo genetic code reprogramming has some potential difficulties, such as unwanted post-translational modifications of the unnatural functional groups present in the npAA[37], and most importantly, *in vivo* there is a strict limit on the degree of reprogramming possible. As the cellular machinery must continue to be made correctly, the pAA’s code cannot be tampered with, and the ‘free’ codons are often limited to the stop codons. Efficient incorporation of more than one npAAs required the use of stop codon reprogramming and 4-base pair codons, together with *E. coli* supplemented with orthogonal translation systems (*i.e.* not just engineered aaRS, but altered ribosomes that recognize a separate Shine Dalgarno sequence)[15c]. Current efforts to alter the genetic code of the organisms themselves, *e.g.* the large scale removal of a chosen stop codon, may help future efforts to ‘free up’ codons for *in vivo* genetic code reprogramming[38].

Flexible tRNA-acylation ribozymes, flexizymes

Genetic code reprogramming can be achieved without using protein catalysed npAA aminoacylation of tRNA. The Suga lab has developed a number of artificial ribozymes ‘flexizymes’ to perform this task[39]. These small, 45 or 46 nucleotides in length, RNA molecules can efficiently catalyse the reaction between *in vitro* transcripts of tRNAs and appropriately activated acyl-donor substrates (Fig. 3A), which can be synthesized easily from commercially available protected pAAs and npAAs[40].

There are three flexizymes currently in routine use and one of these will provide a route to aminoacylated tRNAs for almost any chosen npAAs, be it noncanonical sidechain or backbone chemistry. If the npAA has an aromatic side-chain, in a couple of synthetic steps Boc protected npAA can be activated by cyano-methyl ester, to be recognized by the flexizyme, eFx[40]. If the npAA does not have an aromatic sidechain, it can instead be activated by dinitro-benzyl ester, followed by amino acylation by the flexizyme, dFx[40]. dFx does not recognise the sidechain, rather the dinitro-benzyl ester leaving group itself, providing a general route to attaching any npAA to tRNA regardless of the sidechain. If either of these activated npAA substrates prove to be poorly soluble in the reaction buffer, often due to high hydrophobicity of the sidechain, an alternative activating group, amino-derivatized benzyl thioester, can be used and recognised by the flexizyme aFx[41]. The npAA-tRNA(s) can then be added to a custom-made PURE system, where one or a number of codons are vacant by

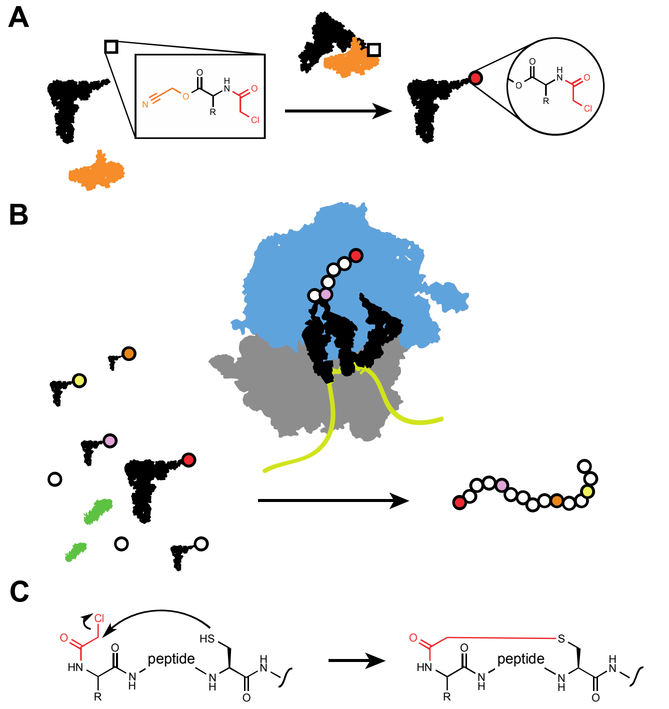


Fig. 3 Ribozyme mediated npAA incorporation. (A) Appropriately activated npAA (white square) can be attached to tRNA (black) using the structured RNA 'flexizyme' (orange). Example shown is an amino acid with unnatural moiety (2-chloroacetyl red) on the amine nitrogen, activated using a CME ester (orange). (B) In a separate reaction, tRNA/npAA conjugate are added to an *in vitro* translation system. In this example the 2-chloroacetyl amino acid is added to the initiator position. Colour scheme identical to Fig. 2. Multiple tRNA/npAA pairs can be prepared separately and added to the same translation mixture for extensive genetic code reprogramming. (C) N-terminal 2-chloroacetyl allows for a nucleophilic cysteine to form the non-reducible linkage that results in a macrocyclic peptide chain.

omitting pAAs and cognate aaRSs, for specific (and multiple) npAA-incorporation into peptides (Fig. 3B). The combination of the flexizyme reaction and the custom-made PURE system allows for enormous freedom with regard to type and number of npAAs incorporated into various peptide sequences.

Fig. 4 shows the enormous variety of npAAs that can be incorporated into elongating peptide chains using the flexizyme methods. These include, unusual side chains, backbone modifications (such as ester bond formation[42], N-methylation[43], and other N-substituted amino acids[44]), and D-stereochemistry[45]. Using flexizymes a number of npAAs can be attached to the tRNA^{fMet} and initiate peptide synthesis (in place of the pAA formylated methionine in *E. coli*). It has been found that this first position in the peptide chain is surprisingly amenable to reprogramming[15b,46]. A large variety of D-stereochemistry pAAs (particularly when the α-amino group is acylated, e.g. with an acetyl group) and even small peptides made up of exotic npAAs[47] can be used as the initiator (Fig. 5).

Of particular importance, in a number of versatile applications, is the incorporation of N-2-chloroacetyl-amino acids at the initiation position[15b] (Fig. 5). This allows for an efficient intramolecular reaction between the C-Cl electrophile and a nucleophilic cysteine further along the peptide chain. The result is a peptide macrocycle, completed by a non-reducible thioether covalent bond[15b] (Fig. 3C). The use of this genetic code reprogramming, in combination with mRNA display (see above), allows for the rapid discovery of peptide macrocycles that bind to biological targets, labelled the RaPID (Random nonstandard Peptides Integrated Discovery) system[8]. The large libraries (trillions of peptides) provided by mRNA display allow for the discovery of macrocyclic peptides that bind tightly (with low to sub nM K_d being typical) and show high selectivity, being even able to discriminate between protein isoforms[48].

With the freedom to include almost any npAA in the RaPID system, N-methyl amino acids were among the first to be included for the protease resistance and increased membrane permeability they can confer. Yamagishi *et al.* heavily reprogrammed the genetic code to include four N-methyl npAA (see Fig 1) and used the RaPID system to select for binders for a ubiquitin ligase E6AP[8], and produced a macrocyclic peptide having a K_d of 0.2 nM. This peptide was able to inhibit the activity of E6AP and was stable towards protease degradation in plasma. Importantly, both binding and plasma stability depended on the npAA mediated macrocyclization and N-methylation present during selection.

Some npAA can be used to introduce a 'warhead' into a peptide library, *i.e.* a rationally chosen functional group designed to interact with a particular target. Using knowledge of the enzyme mechanism, Morimoto *et al.* reprogrammed the genetic code to include ε-*N*-trifluoroacetyl lysine, with the aim to produce potent inhibitors of the human deacetylase SIRT2[49]. They built their DNA libraries to make peptides with this npAA in the centre of a randomized region and, using the RaPID system, a tight binding, ε-*N*-trifluoroacetyl lysine containing, inhibitor was produced.

Another use of the npAA containing peptides produced by the RaPID system is to aid crystallization, and therefore structure determination, of protein targets. Many proteins are difficult to crystallize because of exposed hydrophobic surfaces that may induce aggregation, and the existence of different interconverting conformations[50]. In the presence of RaPID discovered peptides, proteins may be stabilized against precipitation and may be 'locked' into a specific conformation suitable for crystallization. In addition, crystal contacts may be added by the inclusion of the macrocyclic peptide, permitting crystallization. This approach successfully aided the structure determination of a MATE multidrug transporter, a clinically important class of membrane protein[51].

The use of the flexizyme system is arguably the most powerful method of *in vitro* genetic code reprogramming. The use of a particular npAA is not restricted to the recognition by a (wild-type or engineered) aaRS, there is complete freedom over which codons to assign the npAA to and multiple codons can be resigned to a number of npAAs. Yet, a shortcoming of the current technique is that for many difficult to translate amino acids, it is required that wild-type aaRSs must be withdrawn from the translation mix to prevent competition with the pAAs

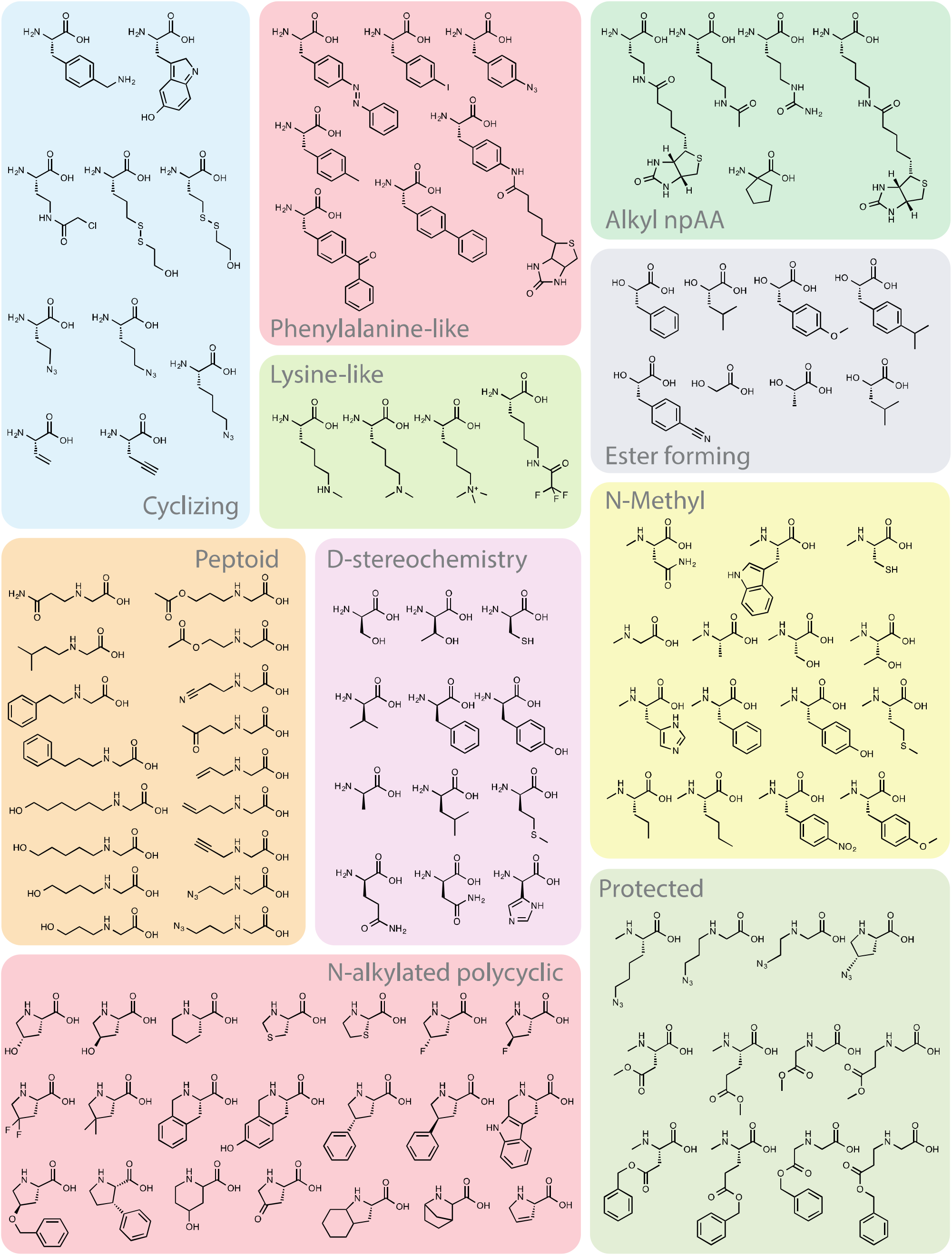


Fig. 4 A collection of npAAs suitable for incorporation at internal positions of peptides, achieved using flexizyme mediated aminoacylation of tRNA and *in vitro* translation. From top left: npAA that allow for macrocyclization (blue)[52]; Phenylalanine-like substrates for eFx (red)[53]; Non-aromatic substrates for dFx (dark green)[40]; Lysine-like substrates (light green) [49]; Substrates for ester bond formation in the ribosome (grey)[42,54]; Peptoids (orange)[44c]. npAA with D-stereochemistry[45] (purple); N-methyl amino acids (yellow)[8,43]; N-alkylated polycyclics (pink)[44a]; Protected npAA allow incorporation charged of N-alkylated amino acids (green)[44b].

(simply withdrawing the potentially competing free amino acid substrate is sometimes insufficient). This means that custom, not currently commercially available, PURE translation systems must be used to achieve the genetic code reprogramming. A less serious shortcoming, which is true for most of the genetic code reprogramming methods described so far, is that vacant codons for the npAA assignment are created by sacrificing pAAs from the genetic code. It would be preferable if all pAAs are still available in addition to the new multiple npAAs in the genetic code.

Non-enzymatic DNA-templated polymerization

While not ‘reprogramming’ any existing the genetic code, the DNA-templated synthesis of macrocyclic peptides pioneered by the Liu lab is highly analogous[55]. Their *in vitro* peptide assembly does not use the natural, ribosomal, translation apparatus. Instead, a template DNA strand, with three artificial

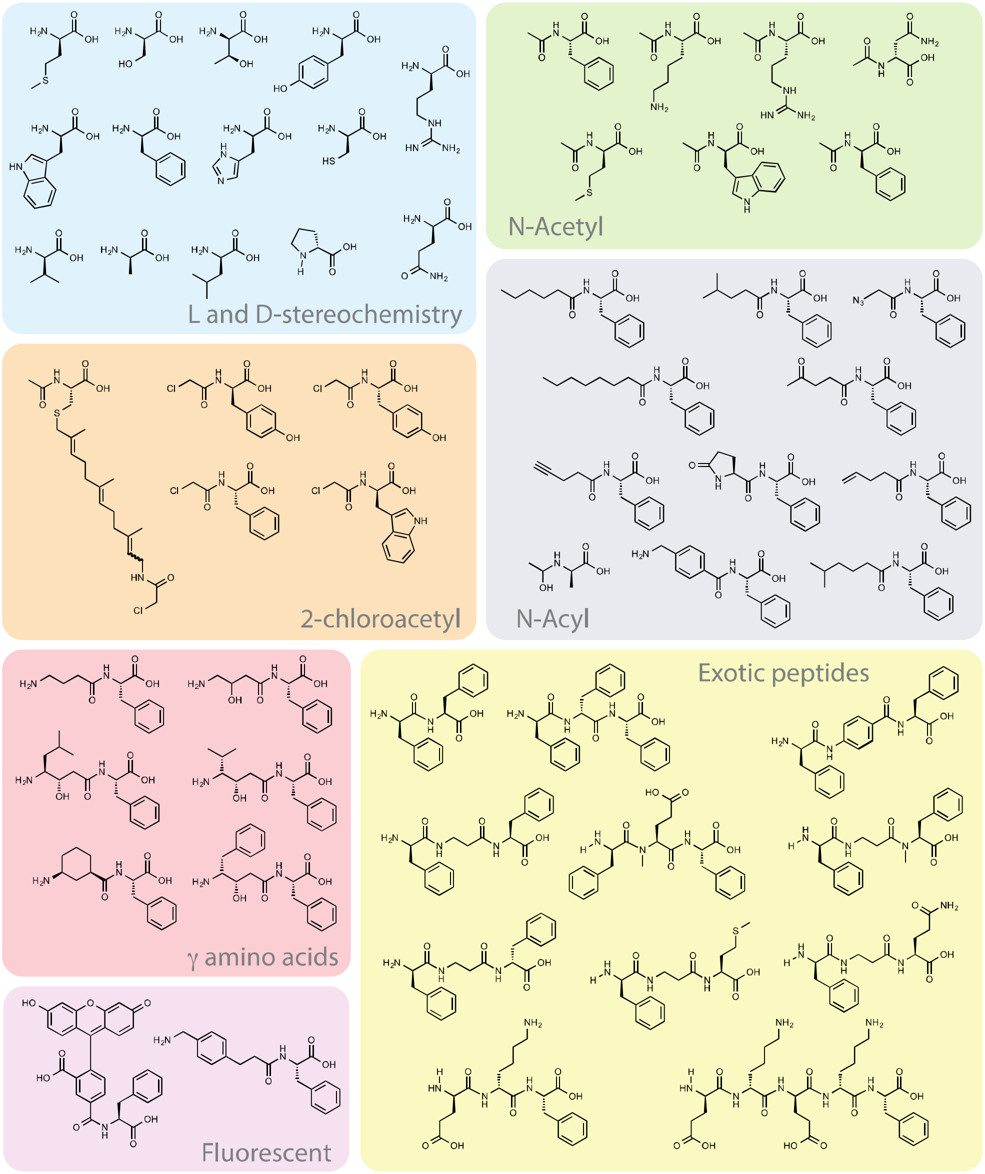


Fig. 5 A collection of npAAs suitable for initiation, achieved using flexizyme mediated aminoacylation of $tRNA^{fMet}$ and *in vitro* translation. From top left: all L-stereochemistry pAA and most D-stereochemistry npAA (blue)[15b,46a46a]; N-Acetyl pAA and npAA (light green)[46b]; 2-chloroacetyl npAA, incorporated for macrocyclization (orange)[8,15b,568,56]; Larger N-Acyl and N-Acetyl moieties (grey); γ-amino acids (red)[57]; Fluorescent, or fluorescent upon reaction (purple)[58]; Exotic peptides (yellow)[47].

'codons' (10-11, as opposed to 3 base pairs) specifying the peptide chain to be formed, is attached to an activated, initiator amino acid[55,59]. Subsequent amino acids are conjugated to a short oligonucleotide, which specifically interacts with the template DNA strand, bringing the amino acid and the growing chain into close proximity, resulting in peptide bond formation. After a final cyclization step, the product is a peptide macrocycle attached to a DNA strand coding for its primary structure – this molecule can be used in a selection process similar to the phage display and mRNA display systems described above. Using this system the Liu lab have found specific binding macrocycles to a range of protein targets[60].

For the ribosome based genetic code reprogramming methods, compatibility with the natural translation machinery places some, still to be fully characterized, limit on how different npAA can be from the pAAs, and on how many, and which, npAAs can be sequentially incorporated[45]. For DNA-templated synthesis, there is no bias towards the natural sidechains, and codes made up of entirely npAA can be generated. Of note, this system is more amenable to backbone modifications, *i.e.* additional $(CH_2)_n$ of β and γ amino acids[59,60b], than ribosomal synthesis[25].

A fundamental difference between the above DNA-directed system and the ribosome-based genetic code reprogramming is that a different codon-anticodon pair must be used for each monomer *and* each position along the chain. Unlike genetic code reprogramming, longer peptide chains require the synthesis of more oligonucleotide linked monomers. This leads to shorter peptides being produced, which drastically reduces the practical library size - in the tens of thousands as opposed to the trillions. However, despite the small size of these macrocycles, and the small library size, the selected molecules can show surprisingly tight binding (K_d/IC$_{50}$ in the μM to 10 nM range).

The future for genetic code reprogramming

There are many directions that genetic code reprogramming could be taken, here we review four areas that are important for the development of functional peptides and proteins.

Pushing the limits of npAA incorporation

A large variety of npAAs have been tested for ribosomal incorporation into peptides, and a large fraction have been suitable substrates for the translation apparatus[7b,25] (see Fig. 4 & 5). However, the chemical space occupied by all the (reasonably sized) npAA is vast[61] and large numbers of npAA remain untested. Further npAAs are likely to be tested in the future and, in the process, our understanding of the preferences and limits of the natural translation system will improved. Related to this, there is currently much effort directed towards modifying the translation machinery to further extend the numbers of npAA suitable for incorporation. In particular, modification to factor EF-Tu, the bodies of tRNAs and the ribosome itself have the potential to facilitate polymerization using challenging monomers[15c,62].

Combining methods of genetic reprogramming

The methods of genetic code reprogramming are not necessarily orthogonal and have the potential to be used in combination. This was recently demonstrated by Ishizawa *et al.*[63] who combined flexizyme-mediated aminoacylation together with promiscuous natural aaRSs to incorporate a number of npAA. This allowed for the incorporation of the initiating 2-choloracetyl-npAA (see above), allowing for macrocyclization, but retained the labour saving incorporation of npAA using existing aaRSs.

Appropriating more from biology

Aside from cyclization, unnatural backbone modification and side-chain stereochemistry there are other aspects of potent bioactive molecules that can be mimicked. One is to reprogram genetic codes to allow peptides to mimic non-peptidic molecules. A recent example showed how initiation could be reprogrammed with a polyketide moiety to mimic amphotericin-B[56]. Another approach is to borrow enzymes used in non-ribosomal peptide synthesis to modify peptides after they have been synthesized by the ribosome. A recent study demonstrated the use of such enzymes to form a number of azoline rings in the backbones of *in vitro* translated peptides[64]. This backbone modification should make these peptides sufficiently different from natural proteins to evade proteolysis and could be a useful tool in functional peptide discovery.

Incorporating structure forming npAA and 'foldamers'

Control over bound, and unbound conformations can have a large effect on the function and behaviour of peptides. As well as affecting the thermodynamics and kinetics of binding, conformational preferences can affect membrane permeability. A striking example is the changes in conformation that enables cyclosporine to cross the cell membrane[65].

Much is known about the structural preferences of individual, natural amino acids[66], and there is a huge body of structural data showing how these can lead to structural motifs and full protein folds. Generally, little is known about the conformational preferences of npAA, and how their inclusion might affect the conformations, dynamics and bound structures of the resulting peptides. Similarly, there is little information about how npAAs interact with the pAAs, or with other npAAs. To address this, a promising future direction is to include npAA(s) intentionally chosen for their predictable conformational preferences. For example, it has long been known that α□aminoisobutyric acid has strong helix inducing properties[67] and that D-stereochemistry amino acids can stabilise some structural motifs[68]. In this direction, some conformational restrained amino acids have been rationally designed to occupy unique ϕ, φ backbone torsional angles [69], and some have been designed to stabilize particular secondary structures or motifs[70]. Future genetic code reprogramming could utilize these new building blocks, and generate codes to include them in optimal combinations. There is likely to be increasing overlap with the field of 'foldamers'[70b,71], where artificial polymers are rationally designed to mimic the ability of biomolecules to fold to defined conformations. While their inclusion may be challenging for genetic code reprogramming, the coupling of rationally designed monomers with the enormous numbers of species tested during selections could be a powerful method of selecting the next generation of functional peptide and peptide-like molecules.

Conclusions

Genetic code reprogramming has allowed a large number of different non-proteinogenic amino acids (npAAs) to be incorporated into peptides. There are three main routes to achieve this reprogramming; utilizing the intrinsic promiscuity of the existing aminoacyl tRNA synthetases (aaRSs), engineering existing aaRSs and using flexizyme catalysed aminoacylation. With regard to the discovery of functional peptides and proteins, each has particular advantages and limitations. If only one npAA needs to be included, or the protein to be discovered (or modified) is large and requires chaperone-assisted folding, *in vivo* methods, using engineered aaRSs, are appropriate. However, when selecting for completely novel proteins and peptides, *in vitro* genetic code reprogramming is generally superior. Using either the promiscuity of aaRSs or flexizyme-mediated aminoacylation allows for a greater degree of reprogramming, *i.e.* multiple reassignments of codons to npAAs. Further, working *in vitro* allows the use of 'mRNA display', permitting the selection of npAA-containing functional peptides from trillions of candidates.

Acknowledgements

This work was supported by the Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST) of Molecular Technologies, the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid Scientific Research (S) (26220204) to H.S., and Grants-in-Aid for JSPS Fellows to J.M.R. (P13766). The authors thank K. Kemplen for critical reading.

Notes and references

1 J. D. Watson and F. H. Crick, *Nature*, 1953, **171**, 737-738.

2 F. H. Crick, L. Barnett, S. Brenner and R. J. Watts-Tobin, *Nature*, 1961, **192**, 1227-1232.

3 Y. Shimizu, A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa and T. Ueda, *Nature Biotechnol.*, 2001, **19**, 751-755.

4 C. R. Woese and G. E. Fox, *Proc. Natl. Acad. Sci. U. S. A.*, 1977, **74**, 5088-5090.

5 L. D. Walensky and G. H. Bird, *J. Med. Chem.*, 2014.

6 W. S. Horne, M. D. Boersma, M. A. Windsor and S. H. Gellman, *Angew. Chem., Int. Ed. Engl.*, 2008, **47**, 2853-2856.

7 (a) H. Neumann, *FEBS Lett.*, 2012, **586**, 2057-2064; (b) C. C. Liu and P. G. Schultz, *Annu. Rev. Biochem.*, 2010, **79**, 413-444.

8 Y. Yamagishi, I. Shoji, S. Miyagawa, T. Kawakami, T. Katoh, Y. Goto and H. Suga, *Chem. Biol.*, 2011, **18**, 1562-1570.

9 S. A. K. Jongkees, C. J. Hipolito, J. M. Rogers and H. Suga, *New J. Chem.*, 2015.

10 V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J. Proteome. Res.*, 2007, **6**, 2351-2366.

11 (a) P. E. Wright and H. J. Dyson, *Curr. Opin. Struct. Biol.*, 2009, **19**, 31-38; (b) J. M. Rogers, V. Oleinikovas, S. L. Shammas, C. T. Wong, D. De Sancho, C. M. Baker and J. Clarke, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 15420-15425.

12 B. Meszaros, P. Tompa, I. Simon and Z. Dosztanyi, *J. Mol. Biol.*, 2007, **372**, 549-561.

13 S. L. Shammas, J. M. Rogers, S. A. Hill and J. Clarke, *Biophys. J.*, 2012, **103**, 2203-2214.

14 J. M. Rogers, C. T. Wong and J. Clarke, *J. Am. Chem. Soc.*, 2014, **136**, 5197-5200.

15 (a) Y. Sako, Y. Goto, H. Murakami and H. Suga, *ACS Chem. Biol.*, 2008, **3**, 241-249; (b) Y. Goto, A. Ohta, Y. Sako, Y. Yamagishi, H. Murakami and H. Suga, *ACS Chem. Biol.*, 2008, **3**, 120-129; (c) H. Neumann, K. Wang, L. Davis, M. Garcia-Alai and J. W. Chin, *Nature*, 2010, **464**, 441-444.

16 D. F. Veber and R. M. Freidinger, *Trends Neurosci.*, 1985, **8**, 392-396.

17 P. K. Madala, J. D. A. Tyndall, T. Nall and D. P. Fairlie, *Chem. Rev.*, 2010, **110**, Pr1-Pr31.

18 (a) D. S. Nielsen, H. N. Hoang, R. J. Lohman, T. A. Hill, A. J. Lucke, D. J. Craik, D. J. Edmonds, D. A. Griffith, C. J. Rotter, R. B. Ruggeri, D. A. Price, S. Liras and D. P. Fairlie, *Angew. Chem., Int. Ed. Engl.*, 2014, **53**, 12059-12063; (b) O. Ovadia, S. Greenberg, J. Chatterjee, B. Laufer, F. Opperer, H. Kessler, C. Gilon and A. Hoffman, *Mol. Pharm.*, 2011, **8**, 479-487; (c) J. Chatterjee, C. Gilon, A. Hoffman and H. Kessler, *Acc. Chem. Res.*, 2008, **41**, 1331-1342; (d) E. Biron, J. Chatterjee, O. Ovadia, D. Langenegger, J. Brueggen, D. Hoyer, H. A. Schmid, R. Jelinek, C. Gilon, A. Hoffman and H. Kessler, *Angew. Chem., Int. Ed. Engl.*, 2008, **47**, 2595-2599; (e) T. R. White, C. M. Renzelman, A. C. Rand, T. Rezai, C. M. McEwen, V. M. Gelev, R. A. Turner, R. G. Linington, S. S. Leung, A. S. Kalgutkar, J. N. Bauman, Y. Zhang, S. Liras, D. A. Price, A. M. Mathiowetz, M. P. Jacobson and R. S. Lokey, *Nat. Chem. Biol.*, 2011, **7**, 810-817.

19 (a) D. Altschuh, O. Vix, B. Rees and J. C. Thierry, *Science*, 1992, **256**, 92-94; (b) F. Giordanetto and J. Kihlberg, *J. Med. Chem.*, 2014, **57**, 278-295.

20 (a) S. Li, S. Millward and R. Roberts, *J. Am. Chem. Soc.*, 2002, **124**, 9972-9973; (b) A. C. Forster, Z. Tan, M. N. Nalam, H. Lin, H. Qu, V. W. Cornish and S. C. Blacklow, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 6353-6357; (c) C. J. Noren, S. J. Anthonycahill, M. C. Griffith and P. G. Schultz, *Science*, 1989, **244**, 182-188.

21 M. Z. Liu, S. Tada, M. Ito, H. Abe and Y. Ito, *Chem. Commun.*, 2012, **48**, 11871-11873.

22 N. Budisa, B. Steipe, P. Demange, C. Eckerskorn, J. Kellermann and R. Huber, *FEBS J.*, 1995, **230**, 788-796.

23 (a) P. Wang, Y. Tang and D. A. Tirrell, *J. Am. Chem. Soc.*, 2003, **125**, 6900-6906; (b) J. K. Montclare, S. Son, G. A. Clark, K. Kumar and D. A. Tirrell, *Chembiochem*, 2009, **10**, 84-86.

24 M. C. Hartman, K. Josephson and J. W. Szostak, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4356-4361.

25 M. C. Hartman, K. Josephson, C. W. Lin and J. W. Szostak, *PLOS One*, 2007, **2**, e972.

26 R. W. Roberts and J. W. Szostak, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 12297-12302.

27 Y. V. G. Schlippe, M. C. T. Hartman, K. Josephson and J. W. Szostak, *J. Am. Chem. Soc.*, 2012, **134**, 10469-10477.

28 F. T. Hofmann, J. W. Szostak and F. P. Seebeck, *J. Am. Chem. Soc.*, 2012, **134**, 8038-8041.

29 S. Greiss and J. W. Chin, *J. Am. Chem. Soc.*, 2011, **133**, 14196-14199.

30 K. Josephson, M. C. Hartman and J. W. Szostak, *J. Am. Chem. Soc.*, 2005, **127**, 11727-11735.

31 F. Tian, M. L. Tsao and P. G. Schultz, *J. Am. Chem. Soc.*, 2004, **126**, 15962-15963.

32 C. C. Liu, A. V. Mack, M. L. Tsao, J. H. Mills, H. S. Lee, H. Choe, M. Farzan, P. G. Schultz and V. V. Smider, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 17688-17693.

33 C. C. Liu, A. V. Mack, E. M. Brustad, J. H. Mills, D. Groff, V. V. Smider and P. G. Schultz, *J. Am. Chem. Soc.*, 2009, **131**, 9616-9617.

34 J. W. Day, C. H. Kim, V. V. Smider and P. G. Schultz, *Bioorg. Med. Chem. Lett.*, 2013, **23**, 2598-2600.

35 M. C. Kang, K. Light, H. W. Ai, W. J. Shen, C. H. Kim, P. R. Chen, H. S. Lee, E. I. Solomon and P. G. Schultz, *Chembiochem*, 2014, **15**, 822-825.

36 T. S. Young, D. D. Young, I. Ahmad, J. M. Louis, S. J. Benkovic and P. G. Schultz, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 11052-11056.

37 J. R. Frost, N. T. Jacob, L. J. Papa, A. E. Owens and R. Fasan, *ACS Chem. Biol.*, 2015.

38 M. J. Lajoie, S. Kosuri, J. A. Mosberg, C. J. Gregg, D. Zhang and G. M. Church, *Science*, 2013, **342**, 361-363.

39 (a) H. Murakami, H. Saito and H. Suga, *Chem. Biol.*, 2003, **10**, 655-662; (b) C. J. Hipolito and H. Suga, *Curr. Opin. Chem. Biol.*, 2012, **16**, 196-203.

40 H. Murakami, A. Ohta, H. Ashigai and H. Suga, *Nat. Methods*, 2006, **3**, 357-359.

41 N. Niwa, Y. Yamagishi, H. Murakami and H. Suga, *Bioorg. Med. Chem. Lett.*, 2009, **19**, 3892-3894.

42 A. Ohta, H. Murakami and H. Suga, *Chembiochem*, 2008, **9**, 2773-2778.

43 T. Kawakami, H. Murakami and H. Suga, *Chem. Biol.*, 2008, **15**, 32-42.

44 (a) T. Kawakami, T. Ishizawa and H. Murakami, *J. Am. Chem. Soc.*, 2013, **135**, 12297-12304; (b) T. Kawakami, T. Sasaki, P. C. Reid and H. Murakami, *Chem. Sci.*, 2014, **5**, 887-893; (c) T. Kawakami, H. Murakami and H. Suga, *J. Am. Chem. Soc.*, 2008, **130**, 16861-16863.

45 T. Fujino, Y. Goto, H. Suga and H. Murakami, *J. Am. Chem. Soc.*, 2013, **135**, 1830-1837.

46 (a) Y. Goto, H. Murakami and H. Suga, *RNA*, 2008, **14**, 1390-1398; (b) Y. Goto, H. Ashigai, Y. Sako, H. Murakami and H. Suga, *Nucleic Acids Symp. Ser.*, 2006, 293-294.

47 Y. Goto and H. Suga, *J. Am. Chem. Soc.*, 2009, **131**, 5040-5041.

48 Y. Hayashi, J. Morimoto and H. Suga, *ACS Chem. Biol.*, 2012, **7**, 607-613.

49 J. Morimoto, Y. Hayashi and H. Suga, *Angew. Chem., Int. Ed. Engl.*, 2012, **51**, 3423-3427.

50 C. J. Hipolito, N. K. Bashiruddin and H. Suga, *Curr. Opin. Struct. Biol.*, 2014, **26**, 24-31.

51 C. J. Hipolito, Y. Tanaka, T. Katoh, O. Nureki and H. Suga, *Molecules*, 2013, **18**, 10514-10530.

52 (a) Y. Sako, J. Morimoto, H. Murakami and H. Suga, *J. Am. Chem. Soc.*, 2008, **130**, 7232-7234; (b) Y. Goto, K. Iwasaki, K. Torikai, H. Murakami and H. Suga, *Chem. Commun.*, 2009, 3419-3421; (c) E. Nakajima, Y. Goto, Y. Sako, H. Murakami and H. Suga, *Chembiochem*, 2009, **10**, 1186-1192.

53 H. Murakami, D. Kourouklis and H. Suga, *Chem. Biol.*, 2003, **10**, 1077-1084.

54 A. Ohta, H. Murakami, E. Higashimura and H. Suga, *Chem. Biol.*, 2007, **14**, 1315-1322.

55 Z. J. Gartner, B. N. Tse, R. Grubina, J. B. Doyon, T. M. Snyder and D. R. Liu, *Science*, 2004, **305**, 1601-1605.

56 K. Torikai and H. Suga, *J. Am. Chem. Soc.*, 2014, **136**, 17359-17361.

57 Y. Ohshiro, E. Nakajima, Y. Goto, S. Fuse, T. Takahashi, T. Doi and H. Suga, *Chembiochem*, 2011, **12**, 1183-1187.

58 (a) Y. Yamagishi, H. Ashigai, Y. Goto, H. Murakami and H. Suga, *Chembiochem*, 2009, **10**, 1469-1472; (b) N. Terasaka, G. Hayashi, T. Katoh and H. Suga, *Nat. Chem. Biol.*, 2014, **10**, 555-557.

59 B. N. Tse, T. M. Snyder, Y. Shen and D. R. Liu, *J. Am. Chem. Soc.*, 2008, **130**, 15611-15626.

60 (a) J. P. Maianti, A. McFedries, Z. H. Foda, R. E. Kleiner, X. Q. Du, M. A. Leissring, W. J. Tang, M. J. Charron, M. A. Seeliger, A. Saghatelian and D. R. Liu, *Nature*, 2014, **511**, 94-98; (b) R. E. Kleiner, C. E. Dumelin, G. C. Tiu, K. Sakurai and D. R. Liu, *J. Am. Chem. Soc.*, 2010, **132**, 11779-11791.

61 M. Meringer, H. J. Cleaves, 2nd and S. J. Freeland, *J. Chem. Inf. Model*, 2013, **53**, 2851-2862.

62 (a) R. Maini, D. T. Nguyen, S. X. Chen, L. M. Dedkova, S. R. Chowdhury, R. Alcala-Torano and S. M. Hecht, *Bioorg. Med. Chem.*, 2013, **21**, 1088-1096; (b) L. M. Dedkova, N. E. Fahmi, S. Y. Golovine and S. M. Hecht, *J. Am. Chem. Soc.*, 2003, **125**, 6616-6617; (c) K. W. Ieong, M. Y. Pavlov, M. Kwiatkowski, M. Ehrenberg and A. C. Forster, *RNA*, 2014, **20**, 632-643.

63 T. Ishizawa, T. Kawakami, P. C. Reid and H. Murakami, *J. Am. Chem. Soc.*, 2013, **135**, 5433-5440.

64 Y. Goto, Y. Ito, Y. Kato, S. Tsunoda and H. Suga, *Chem. Biol.*, 2014, **21**, 766-774.

65 (a) J. G. Beck, J. Chatterjee, B. Laufer, M. U. Kiran, A. O. Frank, S. Neubauer, O. Ovadia, S. Greenberg, C. Gilon, A. Hoffman and H. Kessler, *J. Am. Chem. Soc.*, 2012, **134**, 12125-12133; (b) N. Eltayar, A. E. Mark, P. Vallat, R. M. Brunne, B. Testa and W. F. Vangunsteren, *J. Med. Chem.*, 1993, **36**, 3757-3764; (c) A. Alex, D. S. Millan, M. Perez, F. Wakenhut and G. A. Whitlock, *Medchemcomm*, 2011, **2**, 669-674.

66 (a) L. Serrano and A. R. Fersht, *Nature*, 1989, **342**, 296-299; (b) A. Horovitz, J. M. Matthews and A. R. Fersht, *J. Mol. Biol.*, 1992, **227**, 560-568.

67 K. T. O'Neil and W. F. DeGrado, *Science*, 1990, **250**, 646-651.

68 (a) A. Kumar and V. Ramakrishnan, *Syst. Synth. Biol.*, 2010, **4**, 247-256; (b) A. Rodriguez-Granillo, S. Annavarapu, L. Zhang, R. L. Koder and V. Nanda, *J. Am. Chem. Soc.*, 2011, **133**, 18750-18759.

69 T. T. Tran, H. Treutlein and A. W. Burgess, *Protein Eng. Des. Sel.*, 2006, **19**, 401-408.

70 (a) S. J. Maynard, A. M. Almeida, Y. Yoshimi and S. H. Gellman, *J. Am. Chem. Soc.*, 2014, **136**, 16683-16688; (b) Z. E. Reinert and W. S. Horne, *Org. Biomol. Chem.*, 2014, **12**, 8796-8802.

71 (a) G. Guichard and I. Huc, *Chem. Commun.*, 2011, **47**, 5933-5941; (b) C. M. Goodman, S. Choi, S. Shandler and W. F. DeGrado, *Nat. Chem. Biol.*, 2007, **3**, 252-262; (c) S. H. Gellman, *Acc. Chem. Res.*, 1998, **31**, 173-180.