Polymerase engineering: towards the encoded synthesis of unnatural biopolymers

David Loakes and Philipp Holliger*

Received (in Cambridge, UK) 18th February 2009, Accepted 6th May 2009
First published as an Advance Article on the web 6th July 2009
DOI: 10.1039/b903307f

DNA is not only a repository of genetic information for life, it is also a unique polymer with remarkable properties: it associates according to well-defined rules, it can be assembled into diverse nanostructures of defined geometry, it can be evolved to bind ligands and catalyse chemical reactions and it can serve as a supramolecular scaffold to arrange chemical groups in space. However, its chemical makeup is rather uniform and the physicochemical properties of the four canonical bases only span a narrow range. Much wider chemical diversity is accessible through solid-phase synthesis but oligomers are limited to \( \leq 100 \) nucleotides and variations in chemistry can usually not be replicated and thus are not amenable to evolution. Recent advances in nucleic acid chemistry and polymerase engineering promise to bring the synthesis, replication and ultimately evolution of nucleic acid polymers with greatly expanded chemical diversity within our reach.

1. Polymerase mechanisms of discrimination against non-cognate substrates

DNA polymerases perform an astonishing feat of molecular recognition by incorporating the correct complementary base opposite its template counterpart, with error rates as low as \( 10^{-4} \) in the case of Thermus aquaticus DNA polymerase I (Taq), even in the absence of exonucleolytic proofreading.1 The availability of high-resolution structures for different stages of the polymerase cycle together with elegant work using synthetic base analogues has begun to illuminate the subtle interplay of molecular forces and structural determinants on which the correct readout depends.

Elegant work by Kool and others has shown that formation of the nascent base pair is highly sensitive to geometric aberrations but surprisingly less dependent on Watson–Crick hydrogen bonding (provided other forces, e.g. stacking, can compensate for the loss of binding energy).2 The next step, extension of the incorporated base, appears to depend not only on a correct geometric fit and the formation of a conformationally stable 3'-pair but on the presence of cognate minor groove hydrogen bond donor and acceptor groups. Furthermore, distortions of cognate duplex geometry by the newly formed pair cause significant stalling, not just at the primer 3'-end,3 but up to four bases post-extension.4 Polymerase substrate specificity thus involves passage through three...
different “checkpoints”, involving a readout of distinct molecular parameters, each of which differentially affect incorporation, extension and replication.

This “triple sieve” mechanism poses stringent requirements on the properties of unnatural nucleotide analogues, e.g. hydrophobic base analogues (HBAs) or sugar analogues. While many can be incorporated with good efficiency and specificity due to energetically favourable interactions, the resulting unnatural base pairs are often only very inefficiently extended when at the primer 3’-end or replicated when acting as a templating base.

While the general features of the specificity mechanisms described above apply to most polymerases, the relative importance of the different molecular parameters can vary significantly between different polymerase families and even among members of the same family. For example, polymerases of the polA family, notably Klenow, are able to efficiently form base pairs without Watson–Crick hydrogen bonding (e.g. between difluorotoluene (1) and 4-methylbenzimidazole (2)) provided the nascent base pair displays correct geometry, while the same unnatural bases are utilised with much reduced efficiency by members of the polY family of translesion polymerases.6

A striking example of the sophistication of DNA polymerase molecular discrimination against non-cognate substrates is the exclusion of non-cognate sugar structures. Ribonucleotides differ from deoxyribonucleotides only by a 2'-hydroxyl group on the ribose sugar. The addition of an additional hydroxyl group on the ribose sugar not only renders the product susceptible to degradation by RNases but also alters the structure of DNA, which may interfere with the activity of DNA modifying enzymes. Since the concentration of ribonucleotide triphosphates (NTPs) within the cell (Escherichia coli) is at about 10-fold higher than that of deoxyribonucleotide triphosphates (dTTPs),6 DNA polymerases must be effective at excluding NTPs in order to maintain the integrity of the DNA genome. Indeed, in E. coli DNA polymerase I NTPs are excluded from incorporation by a factor of \(10^3-10^5\)-fold, and this stringent discrimination is mainly encoded in a single residue.7 This “steric gate” residue is positioned such that it occupies the space where the 2'-OH of an incoming NTP would be placed, thereby sterically excluding NTPs from the active site.

2. Incorporation and replication of modified nucleic acids

Great strides have been made in the enzymatic synthesis and replication of nucleic acids comprising unnatural chemical modifications simply through elegant chemistry, judicious use of polymerase and optimisation of reaction conditions and additives. In general, the incorporation and replication of modified nucleotide substrates by polymerases (natural or engineered) comprises a vast body of literature. The work discussed here therefore represents a blend of the personal preferences and interests of the authors and is not meant to be exhaustive.

2.1 Functionalised DNA

The most widely studied group of nucleoside analogues are those comprising base modifications. Substitutions at the C5 position of either dU or dC, placing the substituent in the major groove of the double helix, are especially well tolerated by polymerases and a great variety of C5 substitutions have been explored and found to be compatible with at least sporadic incorporation.

Typically, pyrimidines are modified at the C5 position, and purines at the C7 position (see in particular the extensive works of Seela et al. for the latter). The number of such triphosphate derivatives that have been described are too vast to detail here, but many are dealt with in recent reviews.8-10 The triphosphate derivatives of C5-modified pyrimidines and C7-modified purines are generally very well tolerated by a variety of polymerases (primarily DNA polymerases and reverse transcriptases), thus allowing for the introduction of additional functional groups into nucleic acids for use in, for example, aptamers11,12 and detection.13,14 Purines have also been modified at the C8 position, a modification that generally induces a syn conformation into the nucleoside which has consequences for base pairing. As a result, not all C8-modified purines are good polymerase substrates.15,16

Of particular interest in this group of compounds are fluorescent dye-labelled nucleotides, as they are an essential component of several core technologies of modern biology including sequencing, microarrays and fluorescent in situ hybridisation (FISH). These and other substituted nucleotides provide the opportunity to utilise DNA as a molecular scaffold, i.e. to synthesise nucleic acid polymers displaying substituents at defined positions along the double helix. However, despite systematic efforts to optimise substituent and linker chemistry, many C5-substituted nucleotides, in particular those bearing large hydrophobic dye molecules, remain poor polymerase substrates, presumably due to steric clashes between the large substituents and the polymerase. Efficient replication of such polymers presents further challenges as the bulky substituents in the template strand cause pausing or abortive synthesis.17 Nevertheless, several groups have reported some success with synthesising nucleic acid polymers with several or all of the bases bearing substituents.18-22 The most advanced of these systems was described by Jäger and Famulok, who achieved successful PCR amplification of short (<100 bp) fragments with a number of nucleotide analogues functionalised with small non-fluorescent groups (functionalised DNA (fDNA)) using Pwo polymerase. Such fDNA was found to stain poorly with ethidium bromide and displayed a notable shift in the CD spectrum suggestive of Z-DNA conformation.23,24 Some reports also briefly commented on the poor solubility in aqueous solvents of such polymers.18,23 While optimisation of buffer conditions allowed significant improvements in the synthesis of fDNA, even greater improvements should be achievable through the engineering or evolution of polymerases tailor-made for the synthesis of DNA-polymers bearing substituents.

2.2 Hydrophobic base analogues

A decade ago Kool and co-workers turned the long-held belief of nucleoside-polymerase recognition on its head by demonstrating that certain nucleoside isosteres were
specifically and efficiently incorporated into duplex DNA without Watson–Crick hydrogen bonding.\textsuperscript{25,26} Working with the thymine isostere difluorotoluene (1, X = F) it was shown that \textit{E. coli} DNA polymerase I (Klenow fragment, KF) inserted the \textsuperscript{5}'-triphosphate of (1, X = F) onto the growing primer strand specifically opposite dA, and with an efficiency approaching that of dTTP. Kool argued that hydrogen bond recognition within the polymerase active site was not essential provided there was a good geometric fit for the incoming nucleotide. Kool and co-workers went on to describe the adenine mimics (2, X = CH)\textsuperscript{27} and (2, X = N).\textsuperscript{28} The triphosphates of both of these analogues are inserted opposite dT or (1) by KF specifically and with good efficiency. When these analogues are used as a triphosphate they extend the primer by one nucleotide, whereas when present in the template full-length product can be obtained. The analogue (2, X = N) was used to assess the requirements for minor groove interactions, the nitrogen being equivalent to N3 of adenine. When (2, X = N) is present in the template full-length product is readily achieved with KF, whereas (2, X = CH) is much less efficiently extended, thus demonstrating that minor groove H-bonding interactions do play an important role in polymerase extension.

These analogues also offered an opportunity to explore differences in minor groove H-bonding requirements among different polymerases. Among the polymerases assayed for their ability to bypass these non-polar isosteres of dA and dT, KF(\textit{exo})–, Taq and HIV-RT were all able to extend the (1):(2, X = N) base pair, but not the (1):(2, X = CH) base pair, indicating the requirement for minor groove hydrogen bond recognition. However, Polz, Pol\textbeta{} and T7 DNA polymerases did not extend the non-hydrogen-bonding analogues.\textsuperscript{29,30} NMR studies using \textsuperscript{13}C-methionine-labelled Pol\textbeta{} suggested that the thymine isostere (1) disrupts closure of the polymerase ternary complex, which presumably accounts for why this polymerase is unable to utilise the isostere.\textsuperscript{31}

Single-turnover kinetic analysis with KF(\textit{exo})– found that the presence of (1) decreased catalytic efficiency 30-fold, whilst when paired with adenine isosteres catalytic efficiency is decreased 10\textsuperscript{3}-fold.\textsuperscript{32} As is the case with most unnatural base analogues, incorporation and extension is dependent on polymerases lacking proofreading ability (\textit{exo}). When KF(\textit{exo}+) polymerase is used, it will remove a terminal dT:(2, X = N) base pair with the same efficiency as a mismatch, but interestingly the (1):(2, X = CH) pair is edited much more slowly.\textsuperscript{33} The analogues (1) and (2, X = N) when introduced into \textit{E. coli} as part of a modified phage genome are bypassed with moderate efficiency, and with very high efficiency under SOS-induction. Remarkably, each of the isosteres were read as their equivalent natural base.\textsuperscript{34}

Kool et al. have further explored the geometric fit of thymine isosteres by assessing the polymerase specificity of size analogues of (1) using each of the analogues where X is H, F, Cl, Br and I and their ability to be incorporated opposite dA by KF(\textit{exo}).\textsuperscript{35,36} These analogues present a series in which the base size increases incrementally over the range of 1 Å. Surprisingly, it was the dichloro analogue, which is larger than the natural thymine base, that was incorporated with the greatest efficiency. Similar results were observed with the translesion polymerases Dpo\textalpha{}\textsuperscript{37} and \textit{E. coli} polymerases II and IV,\textsuperscript{38} though in the case of Pol\textalpha{} analogue incorporation efficiency was much reduced. Using both the deoxyribosyl and ribosomal analogues of (1, X = H, F, Cl, Br, I) the DNA- and RNA-dependent synthesis by HIV-RT was explored. DNA synthesis follows the same pattern as described above, whilst with the RNA analogues HIV-RT only discriminates against the smaller analogues.\textsuperscript{39,40}

Isosteric base analogues have also been used to detect the presence of DNA lesions such as abasic sites, which are amongst the most common forms of DNA damage. Matray and Kool et al. demonstrated that the pyrene C-nucleoside (3) will form a specific “base pair” with an abasic site (4) in duplex DNA.\textsuperscript{41,42} The pyrene base occupies a similar space to an entire Watson–Crick base pair (i.e. is an isostere of the missing base pair), as shown by an NMR structure of a DNA duplex containing the (3):(4) pair.\textsuperscript{43} The \textsuperscript{5}'-triphosphate derivative of (3) is specifically incorporated opposite to an abasic site with efficiency similar to that of a natural dNTP by both KF(\textit{exo}), and yeast pol\textalpha{}.\textsuperscript{44} In the latter case, the efficiency of incorporation of pyrene exceeds that of the natural dNTP. It is suggested that the preference for incorporation of pyrene is due to the highly favourable stacking energy of the large aromatic pyrene ring system. However, once pyrene has been incorporated into the primer strand further extension is very inefficient. The large pyrene ring has also been used as a mechanistic probe for DNA polymerase bypass of other DNA lesions. Incorporation of pyrene opposite various template thymine-dimer analogues occurred preferentially opposite the 3'-T of the dimer (except in the case of trans, syn-photodimer) suggesting that, in this case, the 3'-thymine residue is in a conformation that is unfavourable to Watson–Crick base pairing.\textsuperscript{45}

In addition to pyrene, other large aromatic base analogues have been shown to be efficiently incorporated opposite abasic sites by various DNA polymerases. Berdis and co-workers have investigated a number of indole nucleoside derivatives and shown them to be efficiently and specifically incorporated opposite an abasic site by T4 DNA polymerase and KF with up \textasciitilde{}100-fold higher efficiency than dAMP.\textsuperscript{46–49} Once again, once incorporated further extension beyond the indole is not
observed. The most efficient analogues were found to be 5-nitro- and 5-phenyl-indole derivatives,\(^\text{50,51}\) whilst larger\(^\text{52}\) and less planar\(^\text{53}\) derivatives were incorporated less efficiently, as were analogues where \(\pi\)-electron density or \(\pi\)-electron surface area are reduced compared to 5-nitroindole.\(^\text{54}\) All of these support the notion that the energy gain from favourable stacking interactions with a large \(\pi\)-electron surface area can compensate for imperfect shape complementarity in the incorporation step.

### 2.3 New base pairs

In recent years there has been much interest in identifying alternative base-pairing systems. A potential advantage of such a system is that it may allow an expansion of the genetic code, creating new specific codons for the incorporation of unnatural amino acids into proteins. Many different base-pairing systems have been examined. Work in this area was pioneered by Benner et al., who first examined base pairs with unnatural Watson–Crick hydrogen-bonding patterns such as isoguanine (isodG) and isocytosine (isodC).\(^\text{55}\) This modified base pair does not impair the duplex structure, and the nucleotides are substrates for DNA polymerases. However, isodG does exist as tautomers and as such can form base pairs with thymine. The use of 7-deaza(isodG) constrains the analogue into a single tautomeric state and thus improves specificity.\(^\text{56}\) Alternatively, use of 2-thiothymidine, which pairs only inefficiently with the enol tautomer of isodG, leads to efficient use of isodC and isodG in PCR reactions, with 98% retention of the isodC:isodG base pair per PCR cycle.\(^\text{57}\) The analogues isodC and isodG have found application in PCR\(^\text{58}\) and sequencing reactions,\(^\text{59}\) where they are good polymerase substrates, though at slightly reduced efficiency compared with the native nucleotides.

Apart from isodG:isodC Benner has described a family of nucleoside analogues covering many possible permutations of the Watson–Crick hydrogen-bonding patterns, some of which may be suitable as alternative base pairs.\(^\text{60}\) For example, the purine analogue dP (5) forms a specific base pair with the pyrimidine analogue dZ (6), designed to present electron density in the minor groove to facilitate specific polymerase recognition. Each of the triphosphates of dP and dZ are substrates for both A and B family DNA polymerases, forming specific base pairs with better than 95% retention of the dZ:dP base pair per cycle of PCR.\(^\text{61}\)

Hirao et al. have designed a number of unnatural base-pairing systems based on a combination of hydrogen bond recognition and shape complementarity. The base pairs consist of a modified adenine and a pseudobase similar in size to a pyrimidine base. Each base pair occupies about the same space as a cognate base pair, but they form specific base pairs with each other while discriminating against pairing with any of the natural bases due to steric exclusion. For example, the base pair between the 2-thienylpurine (7, \(R = \text{NH}_2, X_1 = \text{CH}\)) and the o xo(1H)pyridine (8, \(X_2 = H\)) forms specifically as there would be a steric clash between either the O4 of thymine or N4 of cytosine and the thiophene ring. In addition, the base pair is further stabilised by the presence of two hydrogen bonds, which also enhance selectivity. These nucleobases form a specific base pair and are a substrate pair for DNA polymerases.\(^\text{62}\) Imidazolin-2-one also serves as a very good partner, with the 2-thienylpurine (7, \(R = \text{NH}_2, X_1 = \text{CH}\)) being a better DNA polymerase substrate than the nucleoside derivative of pyrimid-2-one.\(^\text{63}\) An alternative base pair is one in which the thienyl ring is replaced by a methyl group (7, \(R = H\)) that forms a specific base pair with difluorotoluene (1) but forms a superior base pair, both in terms of oligonucleotide hybridisation and polymerase specificity with the nucleoside derivative of pyrrole-2-carbaldehyde.\(^\text{64,65}\) A further base pair, the imidazo(4,5-b)pyridine derivative of (7, \(R = \text{NH}_2, X_1 = \text{CH}\)), forms a specific base pair with the nucleoside derivative of 2-nitropyrrrole. PCR reactions using the triphosphate derivatives of these analogues resulted in \(\sim 1\%\) of the modified base pair incorporated into the PCR product after just 20 cycles,\(^\text{66}\) and even better efficiency obtained using a pyridoimidazole pairing with a nitropyrrrole (see Fig. 2).\(^\text{67}\)

These analogues are also good substrates for RNA polymerases. Using T7 RNA polymerase the triphosphate derivative of the ribonucleoside of (8, \(X_2 = H\) or I or ariylalkynyl) is efficiently incorporated opposite the thienyl-purine derivatives (7) in a DNA template.\(^\text{68,69}\) Furthermore, the RNA transcripts were substrates for translation using an E. coli cell free system to produce proteins containing 3-chlorotyrosine.\(^\text{70,71}\) Analogues of the ribonucleotide of (8) have been incorporated into RNA transcripts bearing fluorophores,\(^\text{72,73}\) biotin\(^\text{74}\) and iodine useful for photocross-linking to proteins.\(^\text{75}\) Rappaport described an alternative three-base pair set of nucleotides suitable for replication by T7 DNA polymerase.\(^\text{76}\) The three base pairs used are A:T, hypoxanthine:cytosine and 6-thiopurine:5-methyl-2-pyrimidinone. Error frequency rates for these three base pairs were of the order \(10^{-4}-10^{-6}\), with the major mismatches occurring between hypoxanthine and thymine.

Another strategy for orthogonal base pairing through steric exclusion has been explored by Kool et al. They devised a set of nucleotides in which a benzene ring has been inserted between the sugar and the nucleobase, termed xDNA.\(^\text{77,78}\)
and yDNA\textsuperscript{79} (depending on the position of the phenyl ring), and the effect of the additional benzene ring is to widen the duplex structure. Initial studies to determine whether yDNA can transmit genetic information showed that DNA polymerases can insert the correct nucleotide, but at a lower specificity than for normal DNA.\textsuperscript{80}

The group of Ronesberg have investigated a very large range of chemical space in the search for a third base pair based on hydrophobic base analogues. A large body of analogues have been investigated to systematically probe the steric requirements for polymerase recognition, and for formation of homo- and hetero-pairs by the hydrophobic base analogues. The range of chemical space investigated includes substituted phenyl,\textsuperscript{81,82} substituted pyridyl\textsuperscript{83,84} and isocarboxylyl,\textsuperscript{85–87} as well as pyridines,\textsuperscript{88} azaindoles\textsuperscript{89} and other heterocycles.\textsuperscript{90} More recently Ronesberg \textit{et al.} have developed a novel base pair (see Fig. 2) which is replicated by Klenow fragment (\textit{exo–}) DNA polymerases with an efficiency approaching that of a native base pair.\textsuperscript{91}

### 2.4 Universal bases

Universal base analogues\textsuperscript{92,93} are generally characterised as non-hydrogen bonding, hydrophobic base analogues (HBAs) that pair indiscriminately with all the native nucleobases.\textsuperscript{94} Generally they are destabilising in a duplex due to loss of hydrogen bonding and solvation, though this is compensated for by stacking and hydrophobic interactions, the strength of which is related to the size of the HBA. Largely due to their lack of hydrogen bond functionality (but see isosteres above) universal bases are very poorly recognised by polymerases.\textsuperscript{95,96}

However, Ronesberg \textit{et al.} have shown that various isocarboxylyl analogues, such as MICS, behave as efficient universal base analogues, not only in hybridisation terms but also being recognised indiscriminately by KF, albeit at reduced catalytic efficiency.\textsuperscript{85} The universal bases benzimidazole, 5- and 6-nitrobenzimidazole and 5-nitroindole, are incorporated as their 5’-triphosphates opposite native DNA bases by KF and human Polz 4000-fold more efficiently than a mismatch, with Polz preferentially incorporating the universal bases opposite pyrimidines and KF opposite purines.\textsuperscript{97} However, whilst these analogues were shown to bind to Moloney murine leukaemia virus (MMLV)-reverse transcriptase (RT) they were not incorporated by it. Many HBAs can be incorporated into DNA under rather forcing conditions, but once incorporated they are blocking lesions. When present in a template strand dAMP is preferentially incorporated opposite the universal base, with polymerases presumably following the A-rule.\textsuperscript{98}

The exception to this is when incorporated opposite a template lesion, such as an abasic site (see section on isosteres above). The analogue 3-nitropyrrrole has also been shown to be a substrate for the poiovirus RNA polymerase, preferentially incorporating it opposite A and U, though at 100-fold reduced efficiency compared to cognate pairs.\textsuperscript{99}

### 2.5 Modified sugar–phosphate backbone

While there is a growing body of circumstantial evidence to support the idea of a primordial RNA-based genetic system, a plausible and efficient route for the prebiotic synthesis of ribose has been lacking. This has spurred various groups to investigate whether simpler sugar, or sugar-substitutes, can be used to encode genetic information.

Meggere and co-workers have shown that the deoxyribose ring may be substituted for a simple glycerol unit,\textsuperscript{100} GNA (9), which forms stable duplexes with either GNA or (S)-, but not (R)-GNA, with RNA,\textsuperscript{101} and does not form a stable duplex with DNA. Szostack \textit{et al.} have examined a variety of A and B family DNA polymerases as well as reverse transcriptases to determine whether GNA could be replicated as a genetic system. Using (S)-glycerol NTPs it was shown that Therminator DNA polymerase was able to perform +1 extension of a DNA primer–template duplex with an efficiency comparable to incorporation of a native dNTP.\textsuperscript{102} Furthermore, Bst DNA polymerase was able to fully extend across a short region of template GNA, as was an engineered MMLV-RT, albeit to a lesser extent.\textsuperscript{103} The fact that Bst DNA polymerase is able to synthesise DNA from a GNA template suggests that the polymerase is able to sufficiently stabilise the transient DNA-GNA duplex for sequence readout from GNA to be possible.

Another pre-RNA candidate that has been proposed is the mixed acetal aminal of formyl glycerol, FNA, (10). The triphosphate derivatives of FNA have been examined as substrates by DNA polymerases and reverse transcriptases on a DNA template. Unlike GNA, both enantiomers were found to be substrates for a variety of polymerases, notably those with lower fidelity, though the (R)-enantiomer was preferentially incorporated over the ‘natural’ (S)-isomer.\textsuperscript{104}

One of the most intriguing pre-RNA nucleic acids investigated to date involves an aldo sugar containing only four carbon atoms, (1)-\textit{d}–threofuranosyl nucleic acid (TNA), (11), in which the phosphodiester bonds are attached to the 2’- and 3’-positions of the threofuranose ring (3’ → 2’).\textsuperscript{105} TNA duplexes show similar thermal stabilities to that of DNA and RNA, and TNA can also form cross-pairs with either DNA or RNA. X-Ray crystallography has shown that TNA can be readily accommodated into B-form duplex DNA,\textsuperscript{106} but that it is a better mimic for A-form nucleic acids and therefore forms better cross-pairs with RNA.\textsuperscript{107} Cytosine TNA templates specifically direct the non-enzymatic incorporation of activated rGMP leading to RNA products.\textsuperscript{108} Szostak and co-workers have examined in some detail the templating properties of TNA and the incorporation of TNA triphosphates by a variety of DNA polymerases and reverse transcriptases. They have shown that certain polymerases are capable of copying limited regions of template TNA, despite the differences in the sugar–phosphate backbone between DNA and TNA.\textsuperscript{109} Significant improvements in templating ability were observed in the presence of Mn(II) ions. In order to assess the effect of Mn(II) ions on fidelity, polymerase reaction products were sequenced and the error rate compared with template DNA and template TNA. It was shown that using the DNA polymerase Sequenase the error rate for TNA was 1.5-fold higher than for DNA, whilst with the MMLV-RT Superscript II the TNA error rate was 3-fold increased. The authors have suggested that evolution of a TNA-dependent DNA polymerase could lead to improved efficiency, and thus
lead to applications in which TNA was used as genetic information.

\[
\begin{align*}
9 & \quad \text{HO-} & \quad \text{OH} & \quad \text{B} & \quad 10 & \quad \text{HO-} & \quad \text{OH} & \quad \text{B} & \quad 11
\end{align*}
\]

The kinetics of incorporation of TNA triphosphates by DeepVent(exo–) and Bst and HIV-RT have been reported. Whilst the \( k_{\text{cat}} \) and \( K_m \) (or \( V_{\text{max}}/K_m \)) values differ, the trend is the same, which is that the \( K_m \) increases with TNA triphosphates, whilst the \( k_{\text{cat}} \) decreases.\(^{110,111}\) Interestingly, the kinetics of incorporation of a second TNA triphosphate shows a further increase in \( K_m \) and decrease in \( V_{\text{max}}/K_m \).\(^{111}\) However, the polymerases retain fidelity for the incorporation of TNA triphosphates.

The polymerase that has been most widely studied for the incorporation of modified nucleotides and oligonucleotides is Therminator DNA polymerase, an exo–variant of 9′N DNA polymerase, containing an A485L mutation.\(^{112,113}\) It has also been shown to be one of the most effective DNA polymerases for the incorporation of TNA triphosphates. The kinetics of incorporation are much the same as described above for DeepVent exo–DNA polymerase, in that the \( K_m \) is higher and the \( k_{\text{cat}} \) reduced compared to dNTPs. Furthermore, when the primer strand is modified to contain five TNA residues at its 3′-end there is no further significant decrease in efficiency of incorporation of TNA triphosphates.\(^{114}\) Addition of Mn(II) ions further improves the kinetics of incorporation as might be anticipated. Therminator DNA polymerase has been used to generate TNA sequences as long as 80 nt long. Error rates were raised compared to the use of normal dNTPs, but low enough, in principle, for use in SELEX.\(^{115,116}\) The authors suggest that in vitro evolution could be used to identify a mutant polymerase capable of faithful production of TNA oligomers.

A class of modified sugars with interesting properties that has been widely used for SELEX applications are those with 2′-modifications. Among these 2′-F, 2′-NH\(_2\) and especially 2′-OCH\(_3\) have found applications in the aptamer field as they are reasonably well accepted by T7 RNA polymerase and by reverse transcriptases and greatly stabilise aptamers to nuclease degradation.\(^{117,118}\) Indeed, using an optimised protocol of transcription using the Y693F/H784A double mutant of T7 RNA polymerase, Burmeister et al. described the selection of an anti-VEGF aptamer composed entirely of 2′-OCH\(_3\) nucleotides, which was stable in plasma for >96 h and survived autoclaving at 125 °C.\(^{119}\)

Locked nucleic acids (LNA) were first described by Imanishi\(^{120}\) and Wengel,\(^{121}\) and contain a methylene bridge between 2′-O and C4′ of the ribose sugar. The presence of the bridging methylene group locks the sugar into a 3′-endo conformation, reducing the conformational flexibility of the ring. LNA exhibits enhanced binding towards RNA and DNA, and incorporation of LNA into DNA induces A-like conformations. Various DNA polymerases have been assessed for their ability to either bypass LNA within a DNA template, or to incorporate LNA triphosphates. When present in a DNA template LNA is bypassed provided the LNA residues are distributed along the strand.\(^{122}\) The most efficient polymerases were found to be KOD(exo–) and Vent(exo–), though efficiency of replication was reduced compared to native DNA. LNA triphosphates may be incorporated into DNA using Phusion and Therminator DNA polymerases,\(^{123-125}\) again providing there are not too many consecutive substitutions. Pfu exo–DNA polymerase will incorporate LNA triphosphates, but is unable to further extend the product.

Another class of modified sugars that have been studied are those bearing modification at C4′, which have been extensively examined by Marx and co-workers. Such analogues were examined to investigate steric effects in the polymerase active site using various DNA polymerases and reverse transcriptases.\(^{126}\) The effect of a bulky group at C4′ is different between polymerases, and is dependent to some extent on the sequence context. The C4′-methyl derivative is accepted the most readily, although kinetics of incorporation with HIV-RT showed reduced rate of incorporation.\(^{127}\) However, increasing the size of the alkyl substituent to larger than methyl does cause the rate of incorporation to reduce markedly, but increases specificity.\(^{128,129}\)

The final class of sugar-modified nucleoside to be considered here are those based on pyrofuranoose. A study of base-pairing properties of a variety of six-membered carbohydrate mimics with RNA showed that RNA can cross-pair with a broad range of nucleic acid structures.\(^{130}\) Amongst those analogues studied anhydrohexitol nucleic acid (HNA) (12), a homologue of DNA, was found to be the most stable, forming the most stable homo duplex in a polyA-polyT system, as well as forming the most stable cross-pair with RNA (more stable than RNA:RNA). Crystal structures of HNA:HNA\(^{131}\) and of HNA:RNA\(^{132}\) reveal that it adopts an A-form duplex, though with more pronounced major groove, and a more reinforced backbone hydration, which may account for the enhanced stability of such duplexes. HNA oligonucleotides have been used in aptamers,\(^{133}\) and in antisense,\(^{134}\) but are not substrates for RNase H. Various DNA polymerases were able to catalyse the incorporation of a single HNA triphosphate, but only DeepVent(exo–) DNA polymerase could catalyse incorporation of more than one substitution.\(^{135}\) Up to six substitutions could be achieved using DeepVent at high enzyme concentration, though under such conditions fidelity of incorporation was reduced when more than two HNA triphosphates were included in the reaction. The HNA triphosphates have also been tested as substrates for various reverse transcriptases (RTs).\(^{136}\) Whilst all RTs tested were able to incorporate a single HNA triphosphate, only the dimeric RTs RAV2 and HIV-RT were able to incorporate more than one. A series of mutant HIV-RTs were assayed and the M184V mutant of HIV-RT was found to be the most efficient at making multiple incorporations of HNA. HNA incorporated into the AUG start codon and the UUC codon in mRNA did not prevent translation.\(^{137}\) Another interesting six-membered ring replacement for ribose is the cyclohexene nucleoside (CeNA), (13), which like HNA forms exceptionally stable duplexes with
RNA. CeNA has also been used in antisense, but unlike HNA, CeNA oligomers are substrates for RNase H.\textsuperscript{138,139}

3. Polymerase engineering and evolution

As detailed in the previous section, some mutant polymerases show dramatic improvements in their capacity to incorporate, extend and replicate a range of unnatural nucleotide analogues. However, in many cases further improvements would be desirable to enable processive synthesis, replication and evolution of modified nucleic acids. This has provided a strong incentive to engineer polymerase function by design, screening and directed evolution.

Polymerase engineering started with simple deletion of accessory domains of the wild-type polymerase, which sometimes yielded polymerases with improved properties. For example, variants of \textit{E. coli} DNA polymerase I (Klenow) and Taq polymerase (Stoffel fragment and Klentaq) have been generated by full or partial deletion of the N-terminal $5^{\prime}$–$3^{\prime}$ exonuclease domain and show improved stability and fidelity. Mutation of two essential aspartates in the $3^{\prime}$–$5^{\prime}$ exonuclease domain is commonly used to generate proofreading defective variants (e.g. KOD\textsuperscript{exo$^-$}). These \textit{exo$^-$}-variants are often the most useful for the incorporation of unnatural nucleotide substrates, as the proofreading $3^{\prime}$–$5^{\prime}$ exonuclease will efficiently remove most unnatural nucleotides. A particularly interesting mutation turned out to be A485L in the $9^{\prime}$\textit{N} \textit{exo$^-$}-polymerase (Therminator). Originally designed to allow efficient incorporation of acyclic nucleotides for terminator sequencing,\textsuperscript{112,113} it has turned out to improve the incorporation of a wide spectrum of unnatural nucleotide substrates (e.g. TNA, see above).

In some cases careful inspection of high-resolution structures has allowed the rational design of mutants with improved properties. Examples include mutants of Taq polymerase with improved properties of dideoxynucleotide incorporation,\textsuperscript{140} an increased capability to incorporate ribonucleotides,\textsuperscript{141,142} reduced pausing\textsuperscript{143} as well as altered fidelity.\textsuperscript{144} Grafting of new domains has been used to modify polymerase properties, notably processivity. Insertion of the thioredoxin binding loops of T7 or T3 DNA polymerase into the \textit{E. coli} Pol I Klenow fragment or Taq polymerase, respectively, leads to an increase in processivity.\textsuperscript{145,146} Fusion of the \textit{Sulfolobus solfataricus} DNA binding protein Sso7d to Taq and Pfu DNA polymerases also improved processivity in both enzymes.\textsuperscript{147} Insertion of AviTag\textsuperscript{30} peptide loops either side of the primer–template duplex binding cleft of $9^{\prime}$\textit{N} DNA polymerase allowed enzymatic biotinylation and assembly with streptavidin into an artificial “processivity complex”. These complexes tethered to a solid surface lead to an impressive gain in processivity from ca. 20 nt in the unmodified polymerase to several thousand in the modified polymerase.\textsuperscript{148} Loeb et al. pioneered the use of \textit{in vitro} screening by genetic complementation of \textit{E. coli} (polA\textsuperscript{29}, recA718) (Fig. 1a) with Taq polymerase, HIV reverse transcriptase (RT) and human Pol\textbeta.\textsuperscript{149,151} Complementation at non-permissive temperatures was used to probe \textit{inter alia} the mutability of the polymerase active site\textsuperscript{152} and screening of the selected mutants has yielded polymerase variants with a range of properties including an increased capability to incorporate ribonucleotides\textsuperscript{141} or increased fidelity.\textsuperscript{153} Genetic complementation of \textit{Saccharomyces cerevisiae} RAD30, RAD52 has been used to screen variants of human pol\texteta and yielded a pol\texteta mutant with increased activity.\textsuperscript{154}

\textit{In vitro} screening for polymerase function has long been used, for example, in drug discovery notably of antiviral compounds such as nucleotide and non-nucleotide inhibitors of HIV-1 RT.\textsuperscript{155} A number of clever high-throughput assay formats have developed both as semi-quantitative continuous and end-point assays. For example, several groups have used a scintillation proximity assay in which a biotinylated primer is extended with tritium-labelled nucleotide triphosphates and immobilized on streptavidin coated scintillant beads. Other assays take advantage of the increased fluorescence intensity of some intercalating dyes, when bound to the double-stranded DNA (dsDNA) synthesised by an active polymerase.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1}
\caption{Strategies for the selection of polymerase repertoires. (a) Complementation of an \textit{E. coli} strain with a temperature-sensitive mutant of DNA polymerase I (polA$^+$). Only active polymerases (Pol1) support colony growth at the non-permissive temperature (37$^{\circ}$ C). (b) Proximal display of polymerase and primer–template duplex substrate on filamentous phage. Primer extension leads to incorporation of a biotinylated nucleotide which tags the phage displaying the active polymerase (Pol1) for selection. (c) Selection of polymerase function by compartmentalised self-replication within the aqueous droplets of a water-in-oil emulsion (CSR). Only polymerases active under the selection conditions (Pol1) can replicate their own encoding gene and generate “offspring”, while those that are inactive (Pol2) cannot self-replicate and disappear from the gene pool.}
\end{figure}
or detect the incorporation a tagged nucleotide substrate in an ELISA format.\textsuperscript{156} Elegant continuous assays have also been described either based on the change in fluorescence of single-stranded DNA binding displaced upon primer extension or on fluorescence resonance energy transfer (FRET) between acceptor and donor fluorophores located on primer and template strand or as part of a template hairpin structure.\textsuperscript{157} Marx \textit{et al.} have adapted high-throughput liquid-handling to perform \textit{in vitro} screening of polymerase function by assaying duplex DNA synthesis by this hairpin assay\textsuperscript{157} or simply by monitoring the fluorescence increase of the intercalating dye SYBR Green (in primer extension reactions or qPCR). These \textit{in vitro} screening methods were used to identify variants of Taq polymerase with improved mismatch discrimination,\textsuperscript{158} variants of KlenTaq polymerase with substantial RT activity\textsuperscript{159} or with an ability to amplify DNA damaged by UV radiation.\textsuperscript{160}

While \textit{in vivo} screening by genetic complementation is a highly effective and convenient way to identify active variants from a large library of polymerase mutants, its scope is limited by the fact that the only property that is screened for is basal polymerase activity. \textit{In vitro} screening is potentially much more flexible as, in principle, any polymerase property can be optimised, provided it is amenable to a high-throughput assay. On the other hand, most of the assays rely on the detection of double-stranded DNA or on the disruption of template secondary structure by the strand-displacement activity of the polymerase. In many ways, such assays are not ideal for the discovery of polymerases with an enhanced ability to incorporate and replicate unnatural nucleotide substrates as polymerase activity may be too weak for significant amounts of dsDNA to be synthesised or the chemistry of the modified substrates may interfere with dye binding or fluorescence. Other pitfalls include false positives \textit{e.g.} polymerases with an enhanced ability to synthesise dsDNA from just three nucleotides\textsuperscript{160} (\textit{e.g.} avoiding the unnatural substrate), to preferentially misincorporate tagged nucleotides or with an enhanced ability to melt template hairpins. Finally, \textit{in vitro} screening methods are currently limited to polymerase libraries an order of magnitude smaller than those used in genetic complementation.

Larger polymerase repertoires can, in principle, be processed by the use of selection technologies. One of the most productive selection methods has been phage display, in particular for the identification of specific peptide and protein interactions. Both Jestin and Romesberg have adapted phage display for the selection of polymerase activity by proximal display of both primer–template substrate and polymerase on the surface of the phage particle (Fig. 1b).\textsuperscript{161,162} By selection for \textit{in cis} incorporation of a biotin-tagged nucleotide triphosphate into the tethered template–primer duplex Jestin has isolated variants of Taq polymerase with substantial RT activity.\textsuperscript{163} Using a similar approach, Romesberg \textit{et al.} have selected variants of the Stoffel fragment of Taq polymerase with greatly improved incorporation of ribonucleotide triphosphates,\textsuperscript{162} \textit{N}^2\textsuperscript{-}OCH\textsubscript{3}, ribonucleotide triphosphates\textsuperscript{164} and a polymerase with 30-fold improved extension of the unnatural PICS:PICS (14) self-pair.\textsuperscript{165} The phage selection approach should in principle be ideal for the identification of polymerases with enhanced activity with unnatural substrates. The only caveat being that assay conditions must be compatible with phage viability, which limits the selection of polymerases from thermophilic organisms under realistic conditions. Furthermore, the intramolecular tethering of the substrate and polymerase may favour the selection of polymerase variants with low affinity for template–primer duplex and/or poor processivity. Finally, polymerases are usually not efficiently secreted and displayed on the phage surface, although this can be improved by optimisation of the signal peptide.\textsuperscript{166} On the other hand, the phage method should be extremely sensitive and in principle able to detect a single incorporation event.

We have developed an alternative selection strategy for the evolution of polymerases, called “compartmentalised self-replication” (CSR).\textsuperscript{167} CSR is based on a simple feedback loop, in which a polymerase replicates only its own encoding gene with compartmentalisation into the aqueous compartments of a water-in-oil (w/o) emulsion\textsuperscript{168} serving to isolate individual self-replication reactions from each other (Fig. 1c). Thus each polymerase replicates only its own encoding gene to the exclusion of those in other compartments (\textit{i.e.} self-replicates). In such a system adaptive gains directly (and proportionally) translate into genetic amplification of the encoding gene. Therefore, the copy number of a polymerase gene after one round of CSR is correlated to the catalytic activity of the encoded polymerase under the selection conditions, with polymerase genes encoding the most active polymerases best adapted to the selection conditions dominating the population. Segregation of self-replication into discrete, physically separate compartments is critical to ensure linkage of phenotype and genotype during CSR. We developed w/o emulsions that are stable for prolonged periods at temperatures exceeding \(90\) °C and allowing selection of polymerases under a wide range of experimental conditions including PCR thermocycling.\textsuperscript{167} CSR has proved to be a productive method for the selection of polymerase function yielding variants of Taq polymerase with >10-fold increased thermostability,\textsuperscript{167} >130-fold increased resistance to heparin\textsuperscript{167} or a generically enhanced substrate spectrum.\textsuperscript{169} Molecular breeding of polymerase genes from the genus \textit{Thermus} and CSR selection also allowed the isolation of polymerases with a striking ability PCR-amplify damaged DNA, which allowed the increased recovery of DNA sequences from ice-age specimens.\textsuperscript{170} CSR selections from the same polymerase library also yielded a polymerase with a broad resistance to common environmental PCR inhibitors or a generic ability to replicate large hydrophobic base analogues (C. Baar, DL, PH, unpublished results).
CSR makes stringent demands on the catalytic efficiency and processivity of selected polymerases (requiring replication of the >2 kb polymerase gene). In order to reduce the adaptive burden and increase the sensitivity of the method, we devised short-patch CSR (spCSR), in which only a short, defined segment of the polymerase gene is replicated and evolved. spCSR has allowed the isolation of variants of Taq with an expanded substrate range allowing enhanced incorporation of 2'-substituted nucleotides including the generation of mixed RNA–DNA, 2'-F-DNA amplification products in PCR.156 as well as greatly improved incorporation of fluorescent dye-labelled nucleotides171 (N. Ramsay, PH unpublished results).

4. Unnatural biopolymer synthesis

An improved understanding of the molecular mechanisms of polymerase selectivity together with advances in nucleic acid chemistry and polymerase engineering and evolution have brought the synthesis of unnatural nucleic acid polymers within our grasp.

Different types of polymers pose different challenges to substrate design and polymerase engineering.

Nucleic acids can be viewed as repeats of tripartite nucleotide building blocks comprising the nucleobase (A), the sugar scaffold (B) and the backbone linkage (C). Different types of polymers can comprise alternative compounds replacing or modifying either of these components.

To simplify discussion, we propose a classification of unnatural nucleic acid polymers based on this tripartite structure, Fig. 2. Group A polymers DNA (or RNA) polymers comprising standard ribofuranose sugars and phosphoester backbone linkages with modifications to one or all of the four canonical bases. These can be substituents, for example, to the 5'-position of pyrimidines or substitutions of the natural bases with non-canonical ones (as discussed in 2.1–2.4). Group B polymers are DNA (or RNA) polymers comprising either partial or complete replacement of standard ribofuranose sugars by non-standard structures but retaining canonical phosphoester backbone linkages (e.g. GNA, TNA) (as discussed in 2.5). Group C polymers comprise polymers those where the phosphodiester backbone linkages are replaced by alternative structures.

There are many examples of partial group A polymers (in which a natural base is partially replaced by another) synthesised in the laboratory, notably the incorporation of unnatural base pairs.57,62-85-87 Other partial group A polymers are of biotechnological importance as DNA probes used in fluorescent in situ hybridisation (FISH) or in microarrays are generated using fluorescent dye substituted nucleotides. While short probes can be synthesised by a number of polymerases (e.g. Klenow for random prime dye labelling), longer group A polymers can be difficult to synthesise in good yield especially for larger substituents like fluorescent dyes. However, directed evolution by CSR has yielded polymerases that are able to completely replace one of the bases with their fluorescent dye-labelled equivalent (e.g. dA with FITC-12-dA) in PCR.169

Fig. 2 The three panels on the left illustrate the tripartite nucleotide structure, comprising the nucleobase (red), ribofuranose sugar (blue) and phosphoester backbone linkage (green). Nucleobases are either purine bases (A, G) (left) or pyrimidine bases (T, C) (right) and can bear substitutions R1 (yellow) at the X (X = C) or exocyclic amino position (purines) or C5 position (pyrimidines). Panels on the right detail modifications to the canonical structures that are compatible with enzymatic incorporation, such as base substitutions (yellow) (from left to right): alkyne, enamine, alcohol, imidazole, guanidine and pyridyl groups; alternative base pairing systems (red) (from left to right) such as those described by Benner (5,6),61 Yokoyama and Hirao59 and Romesberg;61 sugar modifications (blue) (from left to right), such as hexose nucleic acids (HNA), threose nucleic acids (TNA), glycol nucleic acids (GNA) and 2'-modifications (2'-F, 2'-OCH3) and Locked nucleic acids (LNA); alternative internucleotide linkages (green) (from left to right): methylphosphonates, phosphorothioates, phosphoramidates, boranophosphates and selenophosphates. We propose a classification of nucleic acid polymers, whereby group A polymers comprise nucleic acids with modifications to the base (yellow / red panel), group B polymers comprise modifications to the sugar (blue panel) and group C comprise alternative backbones (green panel).
Complete group A polymers, in which all of the four bases bear a substituent have been described by Famulok for a number of small substituent groups (fDNA). Several other groups have also reported synthesis of group A polymers in which all four bases are substituted with either biotin or a fluorescent dye but with poor yields. Such polymers have been proposed as the basis of a next-generation sequencing method, termed exonuclease sequencing. To be practical more quantitative synthesis of these types of polymers is needed and this is likely to require careful design of substituents to maintain polymer solubility and substantial engineering of the polymerase primer-template binding interface.

Given the stringency of molecular discrimination against non-cognate chemistry by DNA polymerases, one would expect the chemical make-up of natural DNA to be completely uniform. However, the modifications to the cognate DNA chemistry found in nature are diverse and surprisingly wide-spread. They overwhelmingly comprise modifications to the nucleobases (group A polymers) including, apart from the well-known post-replicative epigenetic modifications, the replacement of one of the canonical bases with an unnatural analogue at the level of replication such as the complete replacement of dA by 2,6-diaminopurine in the S. elongatus phage S-2L. These modifications are thought to protect the viral genome from cellular nucleases and promote preferential replication of modified phage DNA by the cognate phage replicases.

Partial and complete group B polymers have so far not been found in nature. However, there are several examples of partial or even complete group B polymers synthesized in the laboratory using unnatural substitutes for ribose (e.g. GNA and TNA) (as discussed in 2.5).

Only one natural example of a group C polymer has been found, that is the zS-modification in Streptomyces. However, completely zS-modified DNA (in which every z oxygen is replaced by sulfur) has been synthesized in the laboratory.

The zS-modification is of course a modest alteration of the backbone linkage have been described and found to be compatible with enzymic synthesis and replication. These include boranophosphate, selenophosphate, phosphoramidate or phosphoamidate linkages. Boranophosphates are good polymerase substrates and have been successfully used in SELEX, and such aptamers may find clinical application in the neuron capture therapy of tumors.

More drastic changes to the nucleic acid structure backbone, in which both phosphoester linkage as well as sugar scaffold are replaced by alternative chemical structures have been attempted but few are satisfactory replacements. A notable exception is PNA in which bases are displayed on a aminomethylglycine backbone. PNA displays highly specific hybridization properties and can direct non-enzymatic replication of complementary DNA. However, enzymatic replication of PNA would require extensive modification to the polymerase active site. It is not inconceivable that polymers such as PNA could be synthesized on an engineered ribosome.

5. Conclusion
Advances in polymerase engineering and evolution together with sophisticated chemistry are poised to enable the enzymatic synthesis, replication and evolution of unnatural nucleic acid polymers with an ever widening array of chemical substitutions. The new sequence and chemical space opened up by these advances is likely to be a rich source of novel nucleic acid therapeutics, aptamers and enzymes with useful applications in medicine, biotechnology, nanotechnology and material science. These novel polymers will provide insights into the parameters of the molecular encoding of information leading to artificial genetic systems based on unnatural chemistry.

Notes and references