DNA polymerases engineered by directed evolution to incorporate non-standard nucleotides

Roberto Laos*, J. Michael Thomson and Steven A. Benner

Foundation for Applied Molecular Evolution, Gainesville, FL, USA

DNA polymerases have evolved for billions of years to accept natural nucleoside triphosphate substrates with high fidelity and to exclude closely related structures, such as the analogous ribonucleoside triphosphates. However, polymerases that can accept unnatural nucleoside triphosphates are desired for many applications in biotechnology. The focus of this review is on non-standard nucleotides that expand the genetic "alphabet." This review focuses on experiments that, by directed evolution, have created variants of DNA polymerases that are better able to accept unnatural nucleotides. In many cases, an analysis of past evolution of these polymerases (as inferred by examining multiple sequence alignments) can help explain some of the mutations delivered by directed evolution.

Keywords: DNA polymerases, non-standard nucleotides, AEGIS, directed evolution, CSR, protein engineering

INTRODUCTION

DNA polymerases are enzymes that catalyze the template-directed synthesis of DNA. Over billions of years, they have evolved to have the speed, specificity, and accuracy required for them to transmit valuable genetic information to and from living organisms with a level of infidelity just sufficient to support Darwinian evolution.

Currently, many DNA polymerases are used in polymerase chain reaction (PCR) and other procedures that involve the copying of nucleic acids. These include multiplexed PCR, nested PCR, reverse transcription PCR, and DNA sequencing. Polymerases are also used to incorporate modified nucleotides, including those that tag, report, or signal the presence of product DNA molecules. They are also now being used to copy sequences built from "artificially expanded genetic alphabets," which add new base pairs to the standard A:T and G:C pair. Together, these technologies are combined to allow nucleic acids to be amplified from complex samples, including saliva, blood, forensic traces, and fossil remains. Furthermore polymerases are supporting *in vitro* selection with expanded genetic alphabets to create receptors that bind to cancer cells (Sefah et al., 2014). Accordingly, the demand for new polymerase variants, especially those with specialized attributes, shows no sign of diminishing, despite the large number of polymerases already available.

This review focuses primarily on polymerase variants that accept nucleic acids having additional nucleotide "letters" that form additional nucleobase pairs. Such expanded genetic systems are being developed in many laboratories (Rappaport, 1988; Switzer et al., 1989; Ishikawa et al., 2000; Tae et al., 2001; Kool, 2002; Geyer et al., 2003; Henry and Romesberg, 2003; Minakawa et al., 2003; Benner, 2004; Hirao et al., 2004; Sismour and Benner, 2005). Some of these simply shuffle the hydrogen bonding groups that join base pairs within a Watson-Crick geometry, such as the artificially expanded genetic information system (AEGIS; Piccirilli et al.,

1990; Geyer et al., 2003). Others attempt to add hydrogen bonds to hold the pair together (Minakawa et al., 2006). Still others hope to dispense with hydrogen bonds entirely (Morales and Kool, 1999). Some polymerases have been modified without the use of directed evolution; however, these cases provide an insight on structure and function of polymerases.

Major advances in "next generation" sequencing, which requires the use of modified nucleotides and DNA polymerases, are considered in a separate review in this series (Chen, 2014).

DNA POLYMERASE FAMILIES

DNA polymerases have been classified into evolutionary families based on an analysis of their amino acid sequences. Initially two decades ago Braithwaite and Ito (1993) used an extensive compilation of the then-available sequences to classify polymerases into three families: A, B, and C. The family names indicated homology to the products of three genes: *pol*A, *pol*B, and *pol*C, which encode for the three canonical polymerases from *Escherichia coli*: DNA polymerase I, DNA polymerase II and DNA polymerase III alpha subunit, respectively (Ito and Braithwaite, 1991; Braithwaite and Ito, 1993). The most studied polymerases belong to Family A (found in prokaryotes, eukaryotes and bacteriophages) and family B (found in prokaryotes, eukaryotes, archaea, and viruses). The D family groups polymerases from Archaea (Cann and Ishino, 1999). Families X and Y are involved in repair. The family X perform base excision repair and double-strand break repair by using their ability to fill short gaps (Moon et al., 2007; Yamtich and Sweasy, 2010). Some polymerases from the family X can perform polymerase activity without template (Berdis, 2014). The family Y groups eukaryotic polymerases (Ohmori et al., 2001) and these show less homology to the previously identified families. Most of family Y polymerases lack proofreading exonuclease domains and have a more open active site to accommodate base damage, presumably this allows them to bypass DNA lesions (Pryor et al., 2014). The RT family groups

the reverse transcriptases, including eukaryotic telomerases and reverse transcriptases found in viruses (Le Grice and Nowotny, 2014).

Early studies recognized that mild proteolysis of DNA polymerase I from *E. coli* produces two fragments, a large fragment that lacks the 5′–3′ exonuclease activity and a small fragment that is then discarded. The large fragment, called the Klenow fragment, retains both the polymerization and proofreading activities of the native enzyme. The Klenow fragment yielded the first crystal structure of a family A polymerase, solved by Ollis et al. (1985). This crystal revealed a "right hand" shape with the active site being located at the "palm" which holds the catalytic amino acids, a "thumb" that binds double-stranded DNA and "fingers" where the incoming nucleotide binds and interacts with the template. The structure of *Thermus aquaticus* DNA polymerase and the analog of the Klenow fragment, the large fragment of *Thermus aquaticus* DNA polymerase (Klentaq1) has also been studied by crystallography (Kim et al., 1995; Korolev et al., 1995).

FAMILY A POLYMERASES

Family A is the most studied of the seven DNA polymerase families. It includes many of the "workhorse" polymerase in classical molecular biology, including the Klenow fragments of *E. coli* and *Bacillus* DNA polymerase I, *Thermus aquaticus* DNA polymerase and the T7 RNA and DNA polymerases. It also includes the first DNA polymerase to be characterized enzymatically, DNA polymerase I from *E. coli*, in seminal work by Kornberg (1960).

Taq polymerase

With the advent of the PCR, it became clear that a polymerase stable to heating would be useful. Here, the DNA polymerase I from *Thermus aquaticus* (*Taq* polymerase) is widely used in PCR. *Thermus aquaticus* was isolated in 1976 from hot springs in Yellowstone National Park (Chien et al., 1976), where it thrives at 70° C. Since the enzyme can be activated by heating the sample and remains active with the high temperatures required to denature DNA strands (typically 94°C), it allows repeated cycles of denaturing, annealing and extension (thermocycling) without the need to add additional polymerase at each cycle. This made PCR a routine laboratory technique.

Eom et al. (1996) solved a co-crystal structure of *Taq* with blunt DNA duplex bound to the active site cleft. This structure had several features: (a) DNA is in an intermediate form between B and A forms. (b) Functionality from certain amino acid side chains hydrogen-bond to the N^3 of purines and the O^2 of pyrimidines of specific residues in the duplex. (c) The 3′ hydroxyl of the primer strand is near three carboxylate groups, delivered by amino acids Asp 785, Glu 786, and Asp610. These are considered to constitute the catalytic core of the enzyme.

As with its homolog, polymerase I from *E. coli*, Taq DNA polymerase can be cleaved to give an active fragment, called Klentaq. This fragment retains polymerase activity without one of its nuclease activities. Li et al. (1998) solved the crystal structures of two ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I (Klentaq1): (a) Klentaq with primer/template and dCTP; (b) Klentaq with primer template. These identified

two conformations of the polymerase: (i) an "open" conformation where the tip of the fingers of the hand is rotated 46° outward and presumably not actively performing the polymerase reaction and (ii) a "closed" conformation, which is "caught in the act" of incorporating a nucleotide. This was the first direct evidence in any DNA polymerase for a large conformational change as part of the catalytic cycle.

Motifs relevant to the engineering of Family A polymerases

Six motifs in the structure of *Taq* Pol are also conserved throughout Family A polymerases. These include motifs A, B, and C (Delarue et al., 1990). Motifs A and B are the most conserved. These two motifs are relevant to DNA polymerase fidelity and substrate specificity, which makes them of special interest to experimentalists seeking to improve the ability of investigators to obtain polymerases that accept unnatural substrates.

Motif A is found in the palm domain of the polymerase and includes the amino acids 605–617; in *Taq*, the sequence is (LLVAL**DY**SQI**E**LR). Within this motif, Asp 610 (bold D) cannot be changed without losing enzymatic activity, presumably, because it coordinates the metal that is directly responsible for catalysis. Glu 615 (bold E) can be changed to Asp without complete loss of activity. Tyr 611(bold Y) which is located in a hydrophobic pocket, can be replaced by a planar aromatic amino acid. The rest of the amino acids in motif A can be replaced by many amino acids without destroying catalytic activity (Patel and Loeb, 2000a).
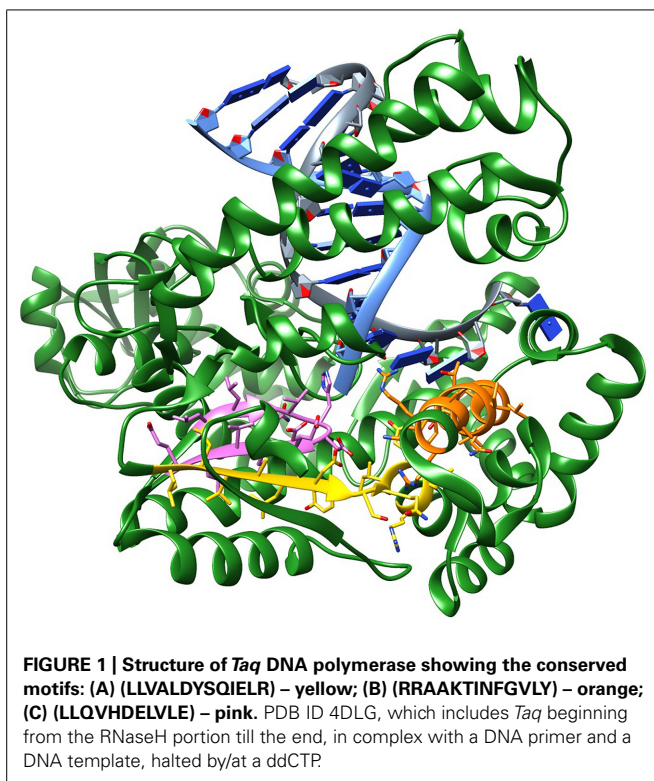
The amino acids that make up motif B are located in the fingers domain and form the O-helix, which contacts the base pair being formed in the primer extension step. This motif covers residues 659–671; in *Taq*, the sequence is (**R**RAA**K**TINFGVLY). This motif contains Arg 659 (bold R) and Lys 663 (bold K), which are known to interact with the incoming triphosphate moiety and are critical for enzymatic activity. For these reasons, they are most likely immutable. Alternatively, Phe 667 and Tyr 671 tolerate conservative substitutions as these are involved in base stacking. The remaining amino acids tolerate a wide range of substitutions. **Figure 1** shows the conserved motifs on the structure of *Taq* polymerase.

Certain substitutions within motif A and B have shown to lower fidelity without eliminating catalytic activity (Suzuki et al., 1997). This is the case of Ile614 in motif A and the Ala661Glu and Thr664Arg substitutions on motif B.

In one example, replacing an amino acid at a single site is known to change substrate specificity in a useful way. Replacement of Phe667 in *Taq* polymerase by a tyrosine eliminates the ability of the polymerase to discriminate against dideoxynucleotides. The *Taq* variant F667Y is, therefore, used for DNA sequencing. Interestingly, in T7 DNA polymerase, the replacement of Tyr526 by Phe increases the discrimination against dideoxynucleotides. This illustrates a general principle in protein engineering: rationales are best constructed *after* the replacement is made and its impact is evaluated.

FAMILY B POLYMERASES

Exploration of the natural microbiosphere led to the discovery of organisms that could grow at temperatures even higher than *Thermus*. These came to be known as "hyperthermophiles,"

FIGURE 1 | Structure of *Taq* DNA polymerase showing the conserved motifs: (A) (LLVALDYSQIELR) – yellow; (B) (RRAAKTINFGVLY) – orange; (C) (LLQVHDELVLE) – pink. PDB ID 4DLG, which includes *Taq* beginning from the RNaseH portion till the end, in complex with a DNA primer and a DNA template, halted by/at a ddCTP.

and were shown by their ribosomal RNA sequences to belong to a third kingdom, or domain, of life of Earth: the Archaea. These proved to be sources of polymerases that were even more thermostable.

For example, *Pyrococcus furiosus* (*Pfu*), a hyperthermophilic archaeon, was discovered in the Lower Geyser Basin of Yellowstone National Park (Brock and Freeze, 1969; Brock and Edwards, 1970). Its DNA polymerase (*Pfu*) has been used in many PCR applications. Crystallographic analysis of the native form (Kim et al., 2008) as well as a variant able to replace dCTP with a cyanine dye-labeled dCTP (Wynne et al., 2013) showed that it contains five distinct domains, called the finger, palm, thumb, N-terminal and exonuclease domains (Hopfner et al., 1999; Hashimoto et al., 2001).

Pyrococcus furiosus has a feature that is absent in *Taq* polymerase: an exonuclease domain that has 3′–5′ exonuclease activity. This allows *Pfu* to proofread using a conformational change (Hopfner et al., 1999). When the polymerase encounters a mismatch, it binds more weakly to the primer/template, causing strand unwinding. This allows the mismatch to move into the exonuclease pocket, where excision ensues (Freemont et al., 1988). A conserved loop in the exonuclease domain interacts with the thumb domain (Kuroita et al., 2005). Mutation of a key residue H147 to a glutamate residue in this loop results in an electrostatic attraction of the thumb domain to the exonuclease domain, preventing the 3′ end of single stranded DNA from entering the exonuclease domain, thus significantly reducing the 3′–5′ exonuclease activity (Wang et al., 1997; Kuroita et al., 2005). Kim et al. (2008) suggested that an alternative residue E148 was located at a better position in the loop to interact with the thumb

domain through a comparison of the crystal structures of *Pfu* and KOD1, a homologous Family B polymerase from the *Thermococcus* genus.

Family B DNA polymerases of hyperthermophilic archaeons may generally have 3′–5′ exonuclease activity (Joyce, 1989; Joyce and Steitz, 1994; Benkovic et al., 2001; Joyce and Benkovic, 2004). Some have shown to recognize the presence of uracil and hypoxanthine in a template strand, stalling when they sense it ahead of the extension site (Greagg et al., 1999; Fogg et al., 2002; Connolly, 2009). This may reflect functional adaptation. Both uracil and hypoxanthine are "mistakes" in a DNA sequence, arising via the deamination of cytosine and adenine, respectively. Such deaminations presumably occur more rapidly at the higher temperatures where hyperthermophiles live. If the polymerase, nevertheless, extends further through the incorporation of dNTPs placing the uracil in the +2 position, the resulting outcome is the activation of the proofreading excision of the deaminated base (Connolly, 2009).

The need for additional proofreading in the natural environment may not be so pressing to a biotechnologist. Accordingly, many have altered the proofreading ability of *P. furiosus* by either removing the exonuclease activity for use in error-prone PCR (ePCR; Biles and Connolly, 2004) or increasing the efficiency of ligation-mediated PCR protocols (Angers et al., 2001). Sanger sequencing also requires the elimination of the exonuclease activity, otherwise incorporated ddNTPs would be removed and the sequencing signals would disappear.

PROTEIN ENGINEERING AND DIRECTED EVOLUTION
THE "NEXT GENERATION" GOALS
For classical Sanger sequencing, a polymerase need only accept a tagged triphosphate with a 3′-blocking group with modest fidelity. The termination:extension ratio will be adjusted in any case by adjusting the concentrations of the terminating and non-terminating triphosphates, meaning that relative inefficiency of incorporation of the unnatural species is not problematic.

However, as the synthetic biology research paradigm has developed, the demands placed on polymerase performance have increased. Here, polymerases are often called upon to copy DNA and PCR amplify molecules containing unnatural nucleotides, often at multiple sites. Here, the fidelity and (preferably) processivity required by a DNA polymerase to support PCR with unnatural nucleotides must be very high. In addition, the structural differences between a DNA polymerase that makes one error per thousand nucleotides and one error per million can be quite subtle and can arise through geometric differences that would not be necessarily distinguished even in a high resolution crystal structure.

RATIONAL DESIGN
Information from structural biology
Molecular biologists would like to believe that they have command of structural theory to "rationally" design polymerases with new, anticipatable properties. In some cases, this is possible, especially when it involves domain shuffling. This has been productive in improving one feature of polymerases important for a wide range of applications: processivity.

DNA binding factors are known to enhance processivity of many polymerases charged with copying complete microbial genomes. In principle, these might be added to improve the processivity of Pol I polymerases, which (as noted above) do not perform this role naturally. Their addition might also, in principle, be used to enhance the performance of any polymerase or polymerase variant. This addition, however, is not often used in biotechnology because of the complexity of the assembled combination. Indeed, *Taq* polymerase and other enzymes are used without accessory proteins for PCR because of their simplicity, which comes from their physiological roles in lagging strand replication and DNA repair.

The complexity of a multicomponent system would be avoided by directly fusing a processivity domain to the active polymerase domain. Adopting this rationale, Wang et al. (2004) covalently fused the double stranded DNA binding protein Sso7d from *Sulfolobus solfataricus* at the N-terminus of *Taq* polymerase (S-*Taq*) and to the fragment of Taq polymerase that results from the deletion of the first 289 amino acids which lacks the exonuclease domain [S-*Taq*(∆289)]. The average length of primer extension prior to template-primer dissociation with *Taq* (∆289) was increased from 2.9 to 51 nucleotides in S-*Taq*(∆289). The full-length *Taq* polymerase, which is intrinsically more processive than *Taq* (∆289), improves its average primer extension from 22 (*Taq*) to 104 (S-*Taq*) nucleotides (Wang et al., 2004).

In parallel work, Wang et al. (2004) also fused the Sso7d domain to the C-terminus of the polymerase from *P. furiosus*, to give *Pfu* polymerase (*Pfu*-S). As in the case of *Taq* polymerase, the fusion of the Ssod7 domain lead to an increase of the average primer extension, from 6.4 nucleotides for *Pfu* to 55 for *Pfu*-S.

Uses for the more processive (*Pfu*-S) were further realized in 1999 when the crystal structure of *Thermococcus gorganarius* DNA polymerase (*Tgo*) was solved. This structure identified a uracil binding pocket, which is used physiologically to prevent the polymerase from copying a template containing uracil, which arises from deamination of cytosine. This structure directed the construction of mutant forms of *Tgo* and *Pfu* DNA polymerases with reduced uracil stalling (Hopfner et al., 1999; Fogg et al., 2002). To increase the ability to read through uracil in the template, the Ssos7 domain was fused to both (*Pfu*-S) and the high fidelity mutant *Pfu* (V93Q). The result was higher processivity and improved uracil-excision cloning (Nour-Eldin et al., 2006).

Structural biology also provided a domain-swapping rational to increase the processivity of *Taq* polymerase. Here, the thioredoxin binding domain (TBD) of the T3 bacteriophage DNA polymerase was inserted into the thumb domain of *Taq* DNA polymerase, deleting amino acids 480–485 (Davidson et al., 2003). The rationale recognized that the processivity of T7 DNA polymerase increases from 15 to 2000 nucleotides when it forms a complex with *E. coli* thioredoxin. The affinity to the primer-template is also increased 80-fold upon binding to thioredoxin. The polymerase arising from this domain fusion remains thermostable, and has a 20–50 times higher processivity than the original *Taq* polymerase.

Exploiting information from multiple sequence alignments (MSAs) in rational engineering

Polymerases are, of course, widely distributed in the biosphere in homologous form. During their divergent evolution, natural selection superimposed upon random variation has carried out several billion years of "protein engineering" experiments, of a sort. With the explosion of microbial sequencing in the last two decades, the results of these "experiments" can be obtained from a public sequence database. To the extent that these results are not corrupted by sequence error, they provide "evolutionary guidance" to assist laboratory protein engineering (Weinhold et al., 1987).

Evolutionary guidance has been productively applied to engineer polymerases, with Tabor and Richardson (1995) providing a classic example. Seeking to improve the ability of *Taq* DNA polymerase I to accept 2′,3′-dideoxynucleoside triphosphates (ddNTPs) for sequencing applications, Tabor and Richardson (1995) examined the sequences of three DNA polymerases from Family A (Braithwaite and Ito, 1993). "Wet" biochemistry had already told them that one of these, that from bacteriophage T7, incorporated ddNTPs better than the two others, polymerases from *E. coli* and *Thermus aquaticus*.

Tabor and Richardson (1995) then constructed a multiple sequence alignment (MSAs) for the three homologous Family A polymerases. They noticed that T7 polymerase had a tyrosine at a site (numbered 526) that is homologous to positions that held a phenylalanine in the *E. coli* and *Taq* polymerases (numbered 762 and 667 respectively). From this comparison, they hypothesized that this single amino acid difference was responsible for the different levels of discrimination against ddNTPs among the three polymerases.

Based on this hypothesis, Tabor and Richardson (1995) replaced the phenylalanine in the *Taq* polymerase by a tyrosine. The result was a variant *Taq* (F667Y) that retained the thermostability of the *Taq* parent but gained improved ability to accept ddNTPs. Similar improvements were seen when the analogous replacement was made in the polymerase from *E. coli*. The mutant *Taq* (F667Y) became one of the first "designed" polymerases to be used in DNA sequencing (Tabor and Richardson, 1995).

Subsequently, Li et al. (1999) studied the crystal structures of Klentaq1, a derivative of *Taq* DNA polymerase that lacks an exonuclease domain. In separate structures, protein crystals binding ddNTPs were observed to have closed ternary complexes, where a conformational change upon substrate binding was associated with a large shift in the position of the side chain of residue 660 in the O helix. Comparing the open and closed structures with ddGTP, Li et al. (1999) concluded that the selective interaction of arginine 660 with the O^6 and N^7 atoms of the G nucleobase might provide structural grounds for better incorporation of ddGTP by *Taq* polymerase. Guided by these observations, Li et al. (1999) then replaced amino acids at residue 660 in Klentaq1 already holding the Tabor-Richardson replacement (F667Y) and studied the resulting variants. Among the variants, the double mutant *Taq* (F667Y; R660D) showed superior performance in DNA sequencing architectures that used ddNTPs.

THE NEED FOR DIRECTED EVOLUTION: SEQUENCE LANDSCAPES

While structure, evolutionary comparison, and mechanistic analysis are all important tools in polymerase engineering, it remains a fact that chemical theory is inadequate to predict the exact outcome of any amino acid replacement on the performance of any protein, including polymerases. A degree of "trial and error" is inherent in protein engineering experiments. This, in turn, requires that we consider the size of the "protein sequence space" that might be explored as we set out to modify a protein to allow it to support a specific technological goal.

Background to this concept was presented by Smith (1970) almost a half century ago. We begin by noting that the behavior of all possible proteins of length n with respect to a measurable behavior can be represented by a space in n dimensions, where each dimension can have one of 20 discrete values, representing the 20 natural amino acids. Each protein sequence is represented by a point in that space. Two points are neighbors in that space if one can be converted into another by a single amino acid substitution. Thus, with 20 amino acids, each point in the sequence space has $19n$ neighbors. The measurable behavior is a real number displayed in the nth $+1$ dimension.

Different sequences have different functions, and moving from a sequence having a function to another functional sequence can proceed via intermediates that either have or lack function. This is illustrated in **Figure 2** with a word game used by Smith (1970), where functional protein sequences are analogous of strings of letters that have a meaning in English. In Smith's (1970) analogy, the sequence of letters in the word "WORD" is converted to the sequence of letters in the word "GENE" by exchanging one letter

at the time, with each step in one path having a meaning (WORE, GORE, and GONE). Paths where all intermediates are meaningful are illustrated by solid lines between points on the surface. Other paths proceed via words lacking meaning, as illustrated by broken lines (for example, WOND, GOND, and GEND).

In this example, linguistic "meaning" is equated to fitness, which provides the nth $+1$ dimension to the surface, a "fitness landscape" (Wright, 1932). The landscape is represented as a topographic map with peaks marked with a $(+)$ for optimal sequences. The absence of function is depicted as dips, marked with a $(-)$. Smith (1970) proposed that natural evolution evolves along paths only if all intermediates are functional. Non-functional sequences are removed by "purifying" selection. Thus, the only valid pathways to explore a sequence space proceed via functional sequences, just as the evolution of words can proceed only via meaningful words.

Sequence space within a protein framework is vast, but enumerable. For example, a 100-amino acid protein can be arranged in 20^{100} different ways. Typical polymerases, eight times longer, constitute a space with 20^{800} points. Both numbers are astronomical. No experiment can sample this space effectively.

Several features of the fitness landscape influence the ease with which it is searched: (a) the fitness landscape is "smooth," meaning that a useful protein sequence can be obtained starting at any point on the landscape via a path that encounters only other functional proteins or, if not, then (b) useful functional proteins can be obtained no matter where one starts the search, as the surface has many of them or, if not, then (c) the library is guided so as to start the search in a region of the functional hypersurface
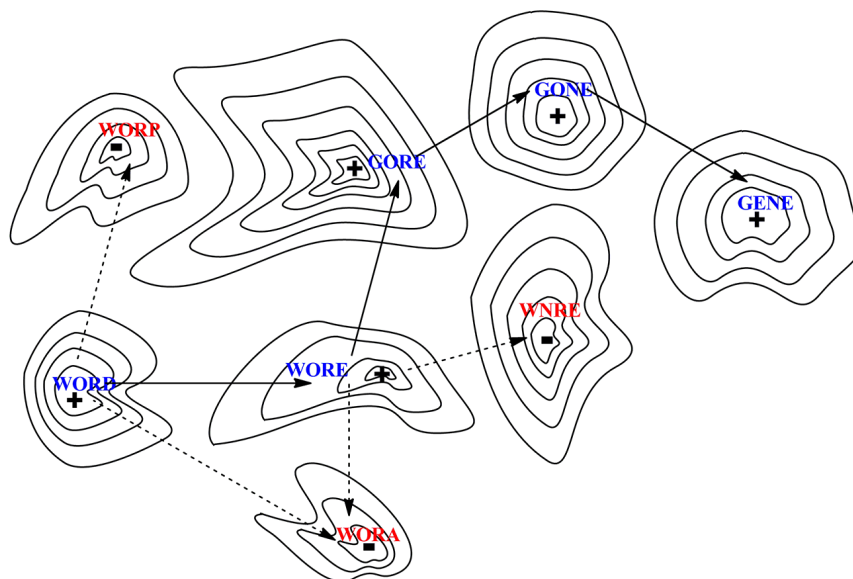
FIGURE 2 | Evolution can be modeled as a walk across a fitness landscape, here presented as a two-dimensional representation of a multiple dimensional hypersurface; analogous to a topographic map, peaks (+) indicate the locations where function exist while dips (–) represent regions with lack of function. Illustrated through an analogy to a word game, a meaningful (functional) string of letters (here "word") must be reached starting from another string ("gene") via stepwise replacement of single letters, where every intermediate along the path must itself also be a functional word. Solid arrows indicate a path of accepted mutations while dashed arrows illustrate deleterious mutations that produce non-functional proteins.

where useful functional proteins reside. Directed evolution is an approach that mimics natural evolution in a time scale that can be reproduced in a laboratory. A directed evolution experiment starts by producing a library of variants (to be discussed further) which then would be selected to a screen or to a selection. A screen involves testing individual variants for the desired properties and is suitable for relatively small libraries, perhaps no more than a few 100s. A selection typically sorts millions of variants at the same time. The experimenter designs the selection in a way that only the variants with the desired properties would survive the selection. The expected outcome of a directed evolution experiment is an enriched pool of variants with proteins having the desired characteristics. Directed evolution can be used to optimize and study any protein (Sterner, 2011).

THE PRACTICE OF DIRECTED EVOLUTION WITH POLYMERASES

In a directed evolution experiment, a "parent" enzyme is chosen to start the search that has (at least) some of the properties desired in the enzyme that will ultimately have utility. The gene of this parent enzyme is then altered to create a library encoding variant forms of the enzyme; some of which might be able to catalyze the desired transformation better than the parent enzyme. The members of the library that are of interest can be isolated by screening or selection.

Library generation

In fact, we have little information about the "smoothness" of any protein fitness landscape. The native polymerase used to initiate an experiment in directed evolution is, of course, already at an elevated point on the fitness landscape, at least for some conditions. It is not clear how many steps (amino acid replacements) can be taken away from the native sequence without losing activity. Further, we expect that certain replacements are more likely to retain core activity than others. All of this suggests that the nature of the library generated from that native sequence might influence the outcome of a directed evolution experiment. It is certainly expected that library generation, if intelligently biased, will allow desired outcomes to be generated faster.

Error prone PCR. A common way to generate libraries from a starting sequence is "mutagenic" or "ePCR." This approach takes advantage of the inherent propensity of *Taq* polymerase to introduce mistakes into the copies of DNA under certain conditions. The frequency of mismatching is often increased by introducing manganese Mn^{2+} along with the natural cofactor Mg^{2+} (Vartanian et al., 1996). Other additives, such as alcohols or unbalanced concentrations of nucleotides, can also be used to introduce mutations through PCR.

The ePCR method produces does not produce a truly random set of amino acid replacements, for several reasons:

(i) *Taq* pol tends to replace purines (adenine and guanine) by other purines and pyrimidines (thymidine and cytidine) by other pyrimidines; these changes are called transitions (as opposed to transversions, which exchange a purine for a pyrimidine or a pyrimidine for a purine). The biased tendency of the polymerase to generate transitions over transversions leads to

libraries with amino acid replacements that are non-random with respect to the parent protein.

(ii) Even if ePCR introduced transitions and transversions equally, the resulting amino acid replacements would not be random, due to the structure of the genetic code. In the code, amino acids having similar chemical properties have closely related codons (Wong et al., 2007). For example, the valine codon (GTN[1]) is converted by a single nucleotide replacement to a phenylalanine codon (TTY[2]), a leucine codon (CTN), an isoleucine codon (ATN), an aspartate codon (GAY) or a glycine codon (GGN). To gain access to codons for other amino acids and, consequently, more dramatically, alter chemical properties in the variant protein, two or three nucleotide replacements are required.

High levels of replacement are not easily achieved by ePCR, nor are they desired. Typical ePCR introduces no more than 4–6 mutations per 1000 nucleotides. Further, a mutation rate that is high enough to search amino acid sequences independent of the code is almost certainly too high to generate any variants that retain polymerase activity as polymerases.

Degenerate codons. Recognizing this challenge, Reetz et al. (2008) developed an elegant approach to library generation that introduces degenerate codons: NNK and NDT. Here, N is any nucleobase, K is guanine or thymine, and D is guanine or adenine or thymine. With the NNK degenerate codon, all 20 amino acids are covered by just 32 ($= 4 \times 4 \times 2$) of the 64 codons possible with standard nucleotides. The twelve NDT degenerate codons ($= 4 \times 3 \times 1$) cover a representative sample of the standard amino acids, including non-polar, aromatic hydrophobic, hydrophilic, and charged amino acids.

Behind this discussion are assumptions about the meaning of the word "random" when discussing amino acid replacements. Some amino acids are encoded by more codons than other amino acids, like serine, with six codons; in contrast, tryptophan is encoded by just one codon. A gene with a truly random sequence would give proteins with a codon-weighed distribution of amino acids. Even this might not be the desired goal of an unguided approach to library generation as some amino acids appear in natural proteins more abundantly than expected from their few codons, for example aspartate and glutamate, each with two codons. Thus, an "ideal" library might arguably be one in which amino acids are replaced by a process that leaves the naturally observed overall composition of the protein unchanged. Finally, our ignorance on the shape of function landscapes, as well as our ignorance of the local topography around any individual parent sequence, means that we cannot state *a priori* which amino acid distribution is most likely to give a desired result in a directed evolution experiment.

Libraries made by gene shuffling or molecular breeding. Random mutagenesis of a parent gene fails, of course, to use all of the information available to a protein engineer, especially in a post-genomic world. As noted above, Nature has already run evolution

---

[1]N is any nucleotide.

[2]Y is a pyrimidine.

experiments. These provide to us many homologs of a parent protein having many amino replacements relative to the parent sequences. Most of these are functional, and, therefore, identify points in sequence space that are elevated on the fitness landscape. It would be desirable to use the information that these homologs provide.

Gene shuffling was introduced by Stemmer (1994) more than a decade ago to directly use these homologs. Here, the starting point is a family of genes that share enough sequence similarity that they can undergo homologous recombination. Using a modified PCR protocol, gene chimeras are produced.

Those using shuffling in protein evolution assume, of course, that sequence space is more efficiently searched by combining the outcomes of two historically successful searches of a particular region of sequence space, than a search that simply replaces single amino acids starting from a single parent. These historical searches delivered the two functioning proteins whose genes are being shuffled. Here, the landscape is assumed to be such that specific paths between two elevated points are also similarly elevated.

This would be a more compelling hypothesis if natural evolution were observed to use shuffling. Natural evolution does, of course, have access to mechanisms that shuffle parts of genes. Natural evolution uses these mechanisms to rearrange (for example) the order of independently folded units in multi-unit polypeptides. This is famously done in the evolution of multi-unit proteins involved in metazoan signal transduction, where a regulatory protein might contain one "src homology domain 1" unit (SH1, a protein kinase), a few SH2 units, and a few SH3 units (Benner et al., 1993). Evolutionary analysis shows that these are all obtained by shuffling, implying that shuffling is an efficient way to search sequence space when no protein folding unit is disrupted.

However, natural evolution does not provide many examples where polypeptide chains within a *single* folded unit are shuffled. This is presumably because the buried contacts binding collections of secondary structural units are finely tuned to permit packing. Changing a single hydrophobic side chain in a packed protein fold often converts a core that is (typically) as densely packed as an organic crystal into a "molten globule." Thus, these biophysical realities would make it surprising to expect that shuffling explores sequence space more effectively than point mutation. Such expectations rely, of course, on the view that natural evolution exploits the most effective ways to search sequence space.

Use of evolutionary information to create smaller but better libraries. Alternative approaches now exist to create libraries that search sequence space around parent sequences (Lutz and Patrick, 2004; Jackel et al., 2008; Lutz, 2010). One class of these exploits evolutionary guidance. For example, Cole and Gaucher (2011) introduced an approach, called the Reconstructing Evolutionary Adaptive Paths (REAP) to create libraries that were hypothesized to explore local sequence space with more efficiency. REAP begins with a phylogenetic analysis of homologous sequences, seeking signatures of functional divergence. An amino acid at a site may be entirely conserved in one branch of a phylogenetic tree, while not conserved at all in a second branch. This pattern of divergence, sometimes called *heterotachy*, indicates that the purifying

selective pressures operating in the first branch at this site are different and stronger than those in the second. This, in turn, means that the function of the proteins within the first phylogenetic branch is different from the function in the second branch.

Only rarely, however, has natural history sought a phenotype desired by a protein engineer, of course, only rarely responsive to the specific adaptive changes needed by today's biotechnologist. Ancient polymerases, for example, were most likely *not* evolving to become resistant to heparin, a target of one of Holliger's selections. Therefore, the rationale for exploiting "evolutionary guidance" is more subtle.

A REAP analysis identifies sites that have been historically involved in *some* adaptive event. Because *some* changes are involved, the amino acid at the site cannot be *absolutely* required for core function. Conversely, the REAP-identified sites are not likely to be those whose amino acids *never* have a phenotypic impact. The rationale being that sites that have in the past been involved in an adaptive event without losing core function are sites that might be productively examined to identify sites that might adapt the protein to the *new*, biotechnologist-demanded, function.

Thus, the rationale behind REAP is the hypothesis that the most productive sites to replace in a protein engineering experiment are neither sites whose amino acids contribute to a core function (as indicated by their absolute conservation) nor sites in which the choice of amino acid is incidental to function (as indicated by their easy variability). By identifying sites for which replacement might have phenotypic impact without destroying core function, REAP is proposed to have an advantage compared to other methods in the generation of libraries with productively altered behaviors. The advantage of the REAP approach relies on the fact that nature has already tested several amino acid sites, and these modifications on these sites produce enzymes that retain the original activity. Searching for new variants in a REAP library gives the advantage of having several parent enzymes.

Thus, the design of a high fidelity DNA polymerase from a medium fidelity polymerase is largely beyond current structure theory. This makes it impossible to get polymerases with the desired high level behaviors from fully guided protein engineering. As a consequence, many investigators use protein engineering to select for polymerases with certain properties improved with respect to a desired function, starting from libraries of polymerase variants. The directed evolution approach is today considered by many to be the method of choice for protein engineering (Bornscheuer and Pohl, 2001; Yuan et al., 2005; Leemhuis et al., 2009; Turner, 2009).

Compartmentalization

Directed evolution requires the connecting of a phenotype with a genotype in a way that allows only genes that confer a desired phenotype to be propagated. This can be done in many ways. One method is compartmentalized self replication (CSR). Developed by Tawfik and Griffiths (1998), CSR holds proteins and genes together in water droplets suspended in oil emulsions. These generally receive the gene-protein pair from a single *E. coli* cell that is encapsulated

within individual droplets (Tawfik and Griffiths, 1998). When the protein is a polymerase variant, its gene is copied only if that variant is active under the conditions of the evolution experiment.

Compartmentalized self replication was first applied to the directed evolution of DNA polymerases by Ghadessy et al. (2001). Here, a library of polymerase genes was delivered in plasmids to create clones in *E. coli* cells. These cells were dispersed into emulsified water droplets containing the primers and buffers needed to perform a PCR amplification of the polymerase gene. Approximately $\sim 10^8 - 10^9$ compartments are formed per milliliter of emulsion; ideally, each compartment contains a single variant. PCR cycling is then performed, with the first heat step lysing the *E. coli* cell to present its expressed thermostable polymerase and its encoding plasmids to the primers. Lysis of the cells then delivers polymerase variants expressed inside of the cells to the buffer, which contains the necessary components for PCR. The polymerase variants and the contents of the buffer remain encapsulated during the PCR cycling.

Polymerases that functioned under the conditions imposed by the experiment were able to make copies of only their own genes. After 20 rounds or more of PCR, the emulsions are broken to give a pool of PCR products enriched in the genes that encoded the selected polymerase variants. These genes could be used directly, or be introduced in cells for another round of
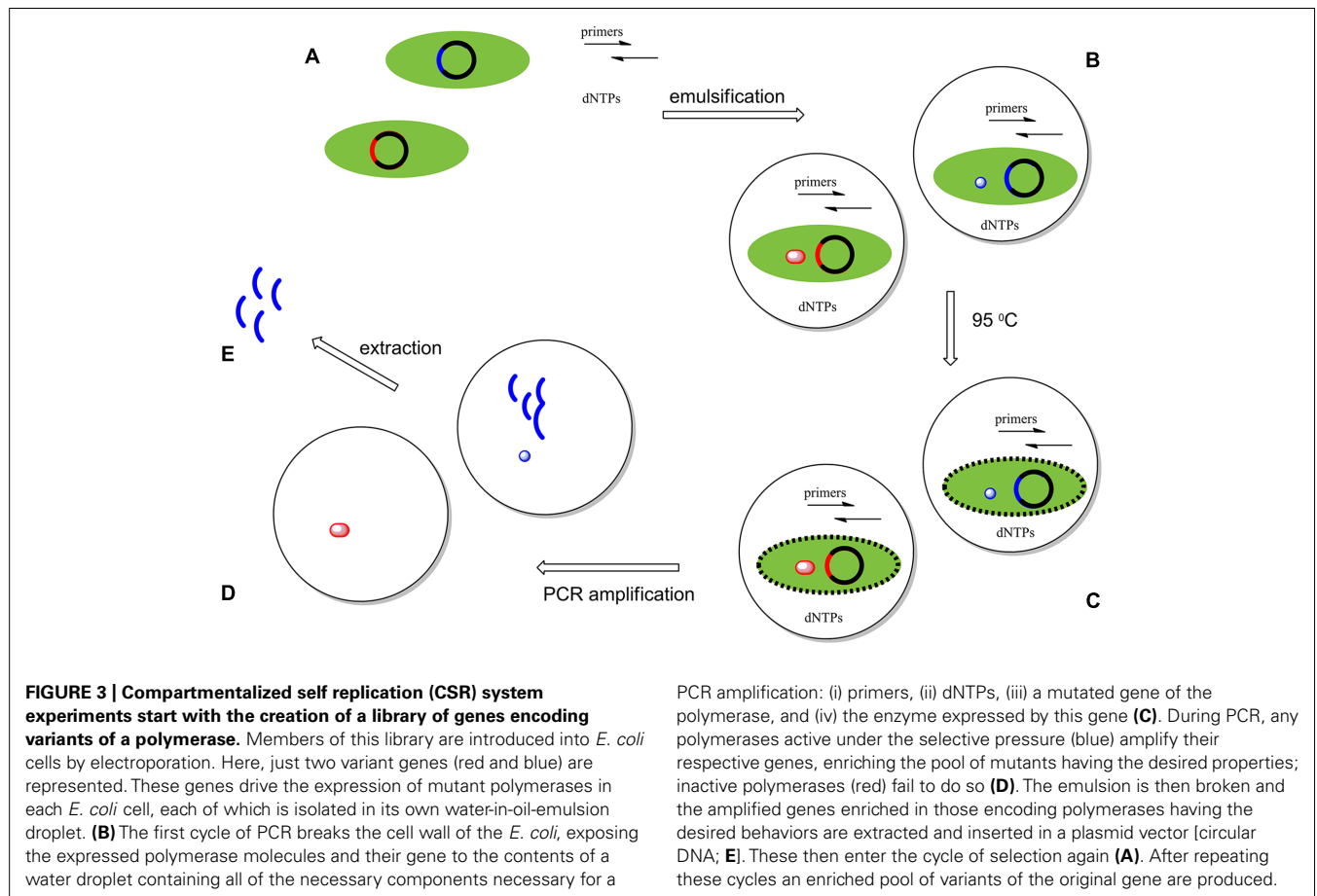
selection. This process is shown schematically in **Figure 3**. With iteration, this process mimics natural evolution, except that the selective pressures applied come from the bioengineer, rather than from Nature.

Phage display is an alternative way to connect genotype and phenotype. In it, a polymerase is linked to its encoding gene in a single viral particle. The protein of interest is co-expressed on the coat of a virus, linking genotype to phenotype. The Romesberg laboratory has been especially active in generating polymerase variants using this approach (Xia et al., 2002; Leconte et al., 2005, 2010).

EXAMPLES OF MODIFIED POLYMERASES
CONVERTING A DNA POLYMERASE TO AN RNA POLYMERASE

Misincorporation by a DNA polymerase through the incorporation of ribonucleoside triphosphates, rather than deoxynucleoside triphosphates, would circumvent the normal pathways in living cells. Accordingly, all DNA polymerases utilize a common mechanism to avoid misincorporation of ribonucleotides by a single active site residue known as the "steric gate" (Joyce, 1997; Gardner and Jack, 1999; Brown and Suo, 2011). Mutations in the steric gate alone are sufficient to render the DNA polymerase able to incorporate nucleoside triphosphates. Yet, products lengths have not exceeded 58 nucleotides and generally result in short termination sequences stalling at +6–7 nucleotides (Gao et al., 1997;

FIGURE 3 | Compartmentalized self replication (CSR) system experiments start with the creation of a library of genes encoding variants of a polymerase. Members of this library are introduced into *E. coli* cells by electroporation. Here, just two variant genes (red and blue) are represented. These genes drive the expression of mutant polymerases in each *E. coli* cell, each of which is isolated in its own water-in-oil-emulsion droplet. **(B)** The first cycle of PCR breaks the cell wall of the *E. coli*, exposing the expressed polymerase molecules and their gene to the contents of a water droplet containing all of the necessary components necessary for a

PCR amplification: (i) primers, (ii) dNTPs, (iii) a mutated gene of the polymerase, and (iv) the enzyme expressed by this gene **(C)**. During PCR, any polymerases active under the selective pressure (blue) amplify their respective genes, enriching the pool of mutants having the desired properties; inactive polymerases (red) fail to do so **(D)**. The emulsion is then broken and the amplified genes enriched in those encoding polymerases having the desired behaviors are extracted and inserted in a plasmid vector [circular DNA; **E**]. These then enter the cycle of selection again **(A)**. After repeating these cycles an enriched pool of variants of the original gene are produced.

Gardner and Jack, 1999; Patel and Loeb, 2000b; Xia et al., 2002; Yang et al., 2002; Ong et al., 2006; McCullum and Chaput, 2009; Brown et al., 2010; Staiger and Marx, 2010; Brown and Suo, 2011). Recently, Cozens et al. (2012) discovered a single amino acid mutation (E664K) in the DNA polymerase from *Thermococcus gorgonarius* that in conjunction with a "steric gate mutation" produced a DNA polymerase capable of synthesizing long RNAs, up 1.7 kb.

Using phage display the Romesberg laboratory has evolved a DNA polymerase [the Stoffel fragment (Sf) of Taq polymerase] into a RNA polymerase. With just five mutations, one of them the "steric gate mutation" (E615G in Taq) the DNA polymerase was able to incorporate ribonucleotides triphosphates (rNTPs) with rates increased by 10^3–10^4 fold compared to the wild type polymerase (Xia et al., 2002). The Holliger laboratory known for the use of the CSR approach has produced a variant of *Taq* polymerase that can incorporate both dNTPs and rNTPs; this variant has only four mutations one of them the "steric gate mutation" mentioned in the previous example (Ong et al., 2006).

DNA POLYMERASES ABLE TO BYPASS DEFECTS

d'Abbadie et al. (2007) shuffled the genes of the polymerases from three *Thermus* species (*aquaticus*, *thermophilus*, and *flavus*) to generate libraries to start a directed evolution experiment to identify DNA polymerases that can extend single, double and quadruple mismatches, process non-canonical primer-template duplexes, and bypass hydantoins and abasic sites (d'Abbadie et al., 2007). They applied these to PCR-amplify cave bear DNA from remains ca. 50 000 years old. These experiments showed that the polymerases obtained by directed evolution applied to these libraries outperformed Taq DNA polymerase and were, therefore, better able to solve a biotechnological problem, here, the sequencing of ancient damaged genomes.

DNA POLYMERASES ABLE TO ACCEPT EXPANDED GENETIC ALPHABETS

One of the AEGIS base pair created in our laboratories is formed between the nucleotides trivially called Z and P (**Figure 4**). The **Z:P** pair has a standard Watson-Crick geometry joined by three hydrogen bonds, differing from the standard C:G pair in the arrangement of donor and acceptor groups that form the connecting hydrogen bonds. Both nucleobases place electron density into the minor groove, a density that can accept a hydrogen bond from a polymerase (Geyer et al., 2003). These features allow polymerases to accept dZTP and dPTP as substrates to form duplexes containing **Z:P** pairs in primer extension reactions, PCR and nested PCR architectures.

In order to improve polymerases able to accept these AEGIS components, we did a selection among Δ(1-279) *Taq* using CSR. Two of the best variants identified: variant (M444V/P527A/D551E/E832V) and variant (N580S/L628V/E832V) showed to pause less when challenged *in vitro* to incorporate dZTP opposite P in a template (Laos et al., 2013). Interestingly, our library was created by introducing random mutations on the *Taq* gene, but the outcome of the selection produced variants which contain several sites that have displayed heterotachy

(different rates of change) in their natural history (Lopez et al., 2002). Heterotachy is a sequence pattern such that the rate of evolution acting at an individual site can be slow in one portion of the phylogeny while the rate at the same site can be rapid in a different portion of the phylogeny. Such patterns arise from shifts in the selective constraints acting at individual sites throughout the evolutionary history of a gene family, and by extension, the precise biomolecular behaviors of the homologous proteins are not identical across the phylogeny (Chen et al., 2010; Cole and Gaucher, 2011) suggesting that these sites were involved in an adaptive change in natural polymerase evolution.

The Romesberg laboratory, using phage display, has produced polymerases having an improved ability to incorporate the self-pairing hydrophobic nucleobases analog propynylisocarbostyril (PICS; Leconte et al., 2005).
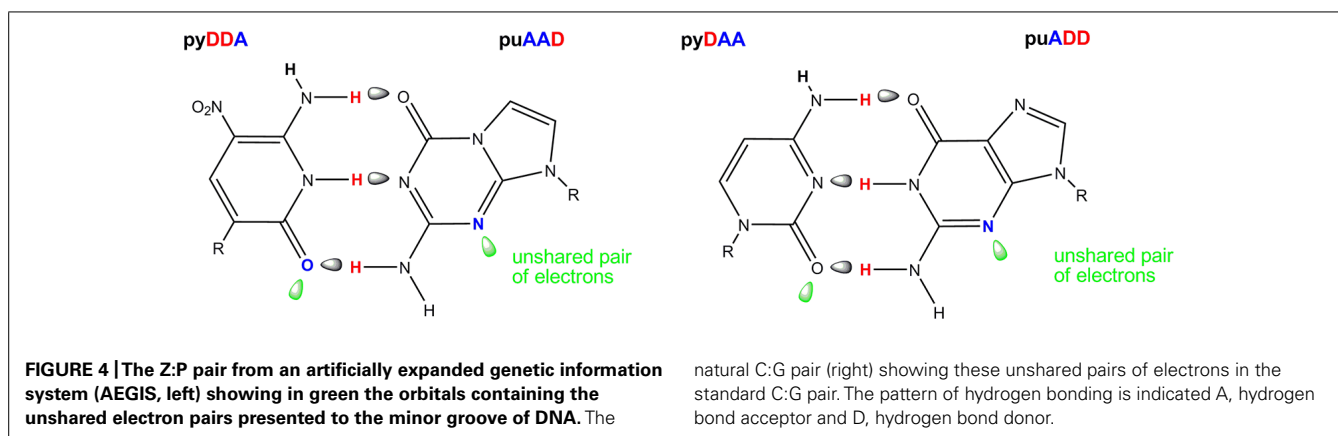
Loakes et al. (2009) have produced a novel polymerase product of the shuffling of polymerases from family A (*Taq* from *T. aquaticus*, *Tth* from *T. thermophilus*, and *Tfl* from *T. flavus*) all of them from the genus *Thermus* and selected by CSR.

POLYMERASES FOR SEQUENCING BY SYNTHESIS METHODS

Sequencing by synthesis (SBS) is a promising next-generation DNA sequencing approach. There are currently several commercial instruments that are offered in the market. Some of the common features of these products are the use of solid phase chemistry to amplify the initial sample and the use of reversible terminators. Reversible terminators are nucleotides that generally have two modifications: one at the 3′OH and the other is either at the 5 or 7 position of the nucleobase. The 3′ hydroxyl position has a cleavable moiety that terminates the polymerase extension reaction after a single-base incorporation. Yet, some reversible terminators do not have a modification on the 3′ hydroxyl like some scarless photocleavable terminator of LaserGen (Wu et al., 2007), the virtual terminator of Helicos BioSciences (Bowers et al., 2009) and the more recently reported Lightning Terminators™ developed in New England Biolabs (Gardner et al., 2012).

The other modification is at the C-5 of pyrimidines or the N-7 position of purines and consist of a fluorescent molecule that is used as a reporter for each of the individual bases. The C-5 and N-7 positions are used because these positions point away from the catalytic pocket of the enzyme. Gardner and Jack (1999) studied variants of Vent DNA polymerase from the hyperthermophilic archaeon *Thermococcus litoralis*. They studied variants on a Tyrosine residue that is highly conserved on family B and was proposed to act as a steric gate (Gardner and Jack, 1999).

The Romesberg laboratories have found polymerases having an improved ability to incorporate modified dUTP with a fluorophore (dUTP-Fl) that can be used for SBS. Leconte et al. (2010) generated a library of Sf, which is *Taq* DNA polymerase minus the first 289 amino acids. This fragment conserves the polymerase activity but lacks the exonuclease domain. The library was done by shuffling the genes of six homologous polymerases: *Thermus aquaticus*; *Thermus thermophilus*; *Thermus caldophilus*; *Thermus filiformis*; *Spirochaeta thermophila*; *and Thermomicrobium roseum*. The three most active polymerase mutants were: Sf168 (with 19 mutations); Sf197 (with 14 mutations). These mutants

FIGURE 4 | The Z:P pair from an artificially expanded genetic information system (AEGIS, left) showing in green the orbitals containing the unshared electron pairs presented to the minor groove of DNA. The natural C:G pair (right) showing these unshared pairs of electrons in the standard C:G pair. The pattern of hydrogen bonding is indicated A, hydrogen bond acceptor and D, hydrogen bond donor.

had between 10 to 50-fold increase in efficiency for dUTP-Fl incorporation compared with wild type Sf (Leconte et al., 2010).

Our laboratories have produced a variant of Taq polymerase using an evolutionary approach to design a polymerase library and then screen a relatively small library for polymerases able to accept unnatural triphosphates modified on their sugar units. Using REAP, they identified 35 sites having heterotachous behavior, after filtering for sites where additional information from evolutionary history, structural biology, and experiments was exploited. They then asked which replacements improve the ability of *Taq* polymerase to accept reversible terminating triphosphates, where the 3′-OH unit of the nucleoside triphosphate had been replaced by an -ONH$_2$ unit, which prevents continued primer extensions. A single modification (L616A) appears to open space behind Phe-667, allowing the enzyme to accommodate a larger 3′-substituent (Chen et al., 2010).

The Holliger lab, when selecting for variants that accepted 2′-deoxycytidine derivatives carrying appended Cy3- and Cy5 fluorescent dyes, recovered variants of *Pfu* DNA polymerase each having two to six amino acid replacements (Ramsay et al., 2010).

COMMERCIAL APPLICATIONS
The number of commercial applications for non-standard nucleotides is large and growing, implying a growing need for engineered polymerases. This review cannot describe all of the potential commercial applications, but a brief summary of those that already exist indicates their scope. For example, Sherrill et al. (2004) used isoC and isoG modified with reporter molecules to develop an assay to detect both RNA and DNA. These modified nucleic acids (isoC and isoG) are also used to quantify levels of HIV and hepatitis viruses in patients (Collins et al., 1997; Elbeik et al., 2004a,b). isoC and isoG are also used to diagnose a panel of respiratory diseases (Nolte et al., 2007).

Sequencing by synthesis technology relies on nucleoside derivatives that are modified in two ways, first with a fluorescent tag, and then (usually) with a reversibly terminating blocking group (Fedurco et al., 2006; Turcatti et al., 2008). Different modified polymerases, which are commercially available have been suggested to improve the procedure (Aird et al., 2011; Fisher et al.,

2011; Quail et al., 2012). Real-time sequencing also requires polymerases that accept nucleoside derivatives (Eid et al., 2009).

At least one polymerase obtained by directed evolution is commercially available; it was selected to the ability to incorporate dZTP opposite dP in a template, and is available through Firebird Biomolecular Sciences LLC (www.firebirdbio.com).

CONCLUSION AND PERSPECTIVES
The demand for polymerases capable of incorporating unnatural nucleotides is certain to grow as the interest to build modified DNA structures continues, including alternative genetic alphabets (Geyer et al., 2003), highly tagged substrates (Hollenstein et al., 2009), modified backbones (Pinheiro et al., 2012), and other unusual structures (Fa et al., 2004; Leconte et al., 2005; Hirao et al., 2007).

The literature teaches that in some cases, simple downstream screening can obtain polymerases with the needed properties. This is illustrated by efforts by Tabor and Richardson (1995), Gardner and Jack (1999), and Chen et al. (2010) to name a few examples. In these cases to create polymerases that accept various 3′-terminating groups. Their combination of structural biology and evolutionary biology analyses were sufficiently powerful to ensure that regions of sequence space small enough to be screened containing polymerases having the desired properties. In each case, screening began with a relatively small number of variants extracted from the sequence space local around a deftly chosen parent, allowing to get useful enzymes by inspection. However, as a result of the large space sequence space of proteins, examples of success in designing polymerases are not frequent, and are rarely (if ever) *de novo*.

In our experience, the outcome of directed evolution experiments can be explained by the analysis of the evolutionary history of the protein; in this case we found the heterotachy pattern. The heterotachy analysis was originally used to elaborate a small library and screen for variants of *Taq* polymerase (Chen et al., 2010). Later we found several of the substitutions recovered in our directed evolution experiment with *Taq* polymerase had this pattern.

It is interesting to note that some mutations reported in the literature by other research groups happen in amino acid sites considered to be heterotachous by our analysis. One particular amino acid change found by our selection of polymerases

better able to synthesize duplexes containing Z:P pairs (D578N) occurred in a site that underwent substitution in the CSR experiment that obtained a *Taq* variant resistant to heparin inhibition (D578G; Ghadessy et al., 2001). Remarkably, position 614 on *Taq* polymerase has been reported at least three times as the outcome of directed evolution experiments (Patel et al., 2001; Xia et al., 2002; Fa et al., 2004). Other amino acid sites from *Taq* polymerase considered heterotachous and reported on directed evolution experiments are: D144, F598 (Ghadessy et al., 2004), A597, A600, E615 (Xia et al., 2002), and L616 (Patel et al., 2001). This provides support for the general hypothesis behind REAP, that sites involved in adaptation to one environmental novelty might also help adaptation to environmental novelties more generally.

We believe that the recapitulation of the natural history of proteins reflects the fact that the sites can be changed to meet new challenges presented to polymerases without damaging the catalytic power or fidelity of the proteins. The observations in the literature underline the importance of understanding the evolution of polymerases in designing libraries to better explore their sequence space. It will be interesting to further study the outcome of contemporary *in vitro* selection experiments and how they recapitulate.

For less effectively constructed libraries of variants, including those generated by shuffling and by undirected mutagenesis, various selection tools stand to pick up where screening cannot possibly go. Here, CSR and phage display have been especially useful. These have yielded polymerases that support the copying of entirely different genetic systems (Pinheiro et al., 2012).

Although some of the mutations found to be useful for altering polymerases to accept unnatural nucleotides fall on or near the conserved motifs and could potentially be rationalized, there are still a number of mutations that cannot be easily explained and their effect could be subtle.

Other approaches have been found to be useful for the evolution of other enzyme systems. For example, neutral drift libraries (Amitai et al., 2007; Bloom et al., 2007a,b), have yet to be applied as starting points for directed evolution experiments with DNA polymerases.

Further, we (and many others) are seeking to develop living systems that implement a "synthetic biology" based on unnatural DNA analogs. These have the potential for being "biosafe" platforms for artificial metabolisms, fermentations, diagnostics, and therapeutic tools, *inter alia* (Schmidt, 2010).

REFERENCES
Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18. doi: 10.1186/gb-2011-12-2-r18

Amitai, G., Gupta, R. D., and Tawfik, D. S. (2007). Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* 1, 67–78. doi: 10.2976/1.2739115

Angers, M., Cloutier, J. F., Castonguay, A., and Drouin, R. (2001). Optimal conditions to use Pfu exo(-) DNA polymerase for highly efficient ligation-mediated polymerase chain reaction protocols. *Nucleic Acids Res.* 29, e83. doi: 10.1093/nar/29.16.e83

Benkovic, S. J., Valentine, A. M., and Salinas, F. (2001). Replisome-mediated DNA replication. *Annu. Rev. Biochem.* 70, 181–208. doi: 10.1146/annurev.biochem.70.1.181

Benner, S. A. (2004). Understanding nucleic acids using synthetic chemistry. *Acc. Chem. Res.* 37, 784–797. doi: 10.1021/ar040004z

Benner, S. A., Cohen, M. A., and Gerloff, D. (1993). Predicted secondary structure for the src homology-3 domain. *J. Mol. Biol.* 229, 295–305. doi: 10.1006/jmbi.1993.1035

Berdis, A. J. (2014). "DNA polymerases that perform template-independent DNA synthesis," in *Nucleic Acid Polymerases*, eds K. S. Murakami and M. A. Trakselis (Berlin: Springer), 109–137.

Biles, B. D., and Connolly, B. A. (2004). Low-fidelity *Pyrococcus furiosus* DNA polymerase mutants useful in error-prone PCR. *Nucleic Acids Res.* 32, e176. doi: 10.1093/nar/gnh174

Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., and Arnold, F. H. (2007a). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.* 5:29. doi: 10.1186/1741-7007-5-29

Bloom, J. D., Romero, P. A., Lu, Z., and Arnold, F. H. (2007b). Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* 2, 17. doi: 10.1186/1745-6150-2-17

Bornscheuer, U. T., and Pohl, M. (2001). Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol* 5, 137–143. doi: 10.1016/s1367-5931(00)00182-184

Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., et al. (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods* 6, 593–595. doi: 10.1038/nmeth.1354

Braithwaite, D. K., and Ito, J. (1993). Compilation, alignment, and phylogenetic-relationships of dna-polymerases. *Nucleic Acids Res.* 21, 787–802. doi: 10.1093/nar/21.4.787

Brock, T. D., and Edwards, M. R. (1970). Fine structure of *Thermus aquaticus*, An Extreme thermophile. *J. Bacteriol.* 104, 509–517. doi: 10.1073/pnas.0712334105

Brock, T. D., and Freeze, H. (1969). *Thermus aquaticus* gen N and sp N A nonsporulating extreme thermophile. *J. Bacteriol.* 98, 289–297.

Brown, J. A., Fiala, K. A., Fowler, J. D., Sherrer, S. M., Newmister, S. A., Duym, W. W., et al. (2010). A novel mechanism of sugar selection utilized by a human X-family DNA polymerase. *J. Mol. Biol.* 395, 282–290. doi: 10.1016/j.jmb.2009.11.003

Brown, J. A., and Suo, Z. (2011). Unlocking the sugar "Steric Gate" of DNA polymerases. *Biochemistry* 50, 1135–1142. doi: 10.1021/bi101915z

Cann, I. K. O., and Ishino, Y. (1999). Archaeal DNA replication: identifying the pieces to solve a puzzle. *Genetics* 152, 1249–1267.

Chen, C. Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. *Front. Microbiol.* 5:305. doi: 10.3389/fmicb.2014.00305

Chen, F., Gaucher, E. A., Leal, N. A., Hutter, D., Havemann, S. A., Govindarajan, S., et al. (2010). Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1948–1953. doi: 10.1073/pnas.0908463107

Chien, A., Edgar, D. B., and Trela, J. M. (1976). Deoxyribonucleic-acid polymerase from extreme thermophile *Thermus-aquaticus*. *J. Bacteriol.* 127, 1550–1557.

Cole, M. F., and Gaucher, E. A. (2011). Exploiting models of molecular evolution to efficiently direct protein engineering. *J. Mol. Evol.* 72, 193–203. doi: 10.1007/s00239-010-9415-9412

Collins, M. L., Irvine, B., Tyner, D., Fine, E., Zayati, C., Chang, C. A., et al. (1997). A branched DNA signal amplification assay for quantification of nucleic acid targets below 100 molecules/ml. *Nucleic Acids Res.* 25, 2979–2984. doi: 10.1093/nar/25.15.2979

Connolly, B. A. (2009). Recognition of deaminated bases by archaeal family-B DNA polymerases. *Biochem. Soc. Trans.* 37, 65–68. doi: 10.1042/bst0370065

Cozens, C., Pinheiro, V. B., Vaisman, A., Woodgate, R., and Holliger, P. (2012). A short adaptive path from DNA to RNA polymerases. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8067–8072. doi: 10.1073/pnas.1120964109

d'Abbadie, M., Hofreiter, M., Vaisman, A., Loakes, D., Gasparutto, D., Cadet, J., et al. (2007). Molecular breeding of polymerases for amplification of ancient DNA. *Nat. Biotechnol.* 25, 939–943. doi: 10.1038/nbt1321

Davidson, J. F., Fox, R., Harris, D. D., Lyons-Abbott, S., and Loeb, L. A. (2003). Insertion of the T3 DNA polymerase thioredoxin binding domain enhances the processivity and fidelity of Taq DNA polymerase. *Nucleic Acids Res.* 31, 4702–4709. doi: 10.1093/nar/gkg667

Delarue, M., Poch, O., Tordo, N., Moras, D., and Argos, P. (1990). An attempt to unify the structure of polymerases. *Protein Eng.* 3, 461–467. doi: 10.1093/protein/3.6.461

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi: 10.1126/science.1162986

Elbeik, T., Markowitz, N., Nassos, P., Kumar, U., Beringer, S., Haller, B., et al. (2004a). Simultaneous runs of the bayer VERSANT HIV-1 version 3.0 and HCV bDNA version 3.0 quantitative assays on the system 340 platform provide reliable quantitation and improved work flow. *J. Clin. Microbiol.* 42, 3120–3127. doi: 10.1128/jcm.42.7.3120-3127.2004

Elbeik, T., Surtihadi, J., Destree, M., Gorlin, J., Holodniy, M., Jortani, S. A., et al. (2004b). Multicenter evaluation of the performance characteristics of the Bayer VERSANT HCV RNA 3.0 assay (bDNA). *J. Clin. Microbiol.* 42, 563–569. doi: 10.1128/jcm.42.2.563-569.2004

Eom, S. H., Wang, J. M., and Steitz, T. A. (1996). Structure of *Taq* polymerase with DNA at the polymerase active site. *Nature* 382, 278–281. doi: 10.1038/382278a0

Fa, M., Radeghieri, A., Henry, A. A., and Romesberg, F. E. (2004). Expanding the substrate repertoire of a DNA polymerase by directed evolution. *J. Am. Chem. Soc.* 126, 1748–1754. doi: 10.1021/ja038525p

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* 34, e22. doi: 10.1093/nar/gnj023

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., et al. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* 12, R1. doi: 10.1186/gb-2011-12-1-r1

Fogg, M. J., Pearl, L. H., and Connolly, B. A. (2002). Structural basis for uracil recognition by archaeal family B DNA polymerases. *Nat. Struct. Biol.* 9, 922–927. doi: 10.1038/nsb867

Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R., and Steitz, T. A. (1988). Cocrystal structure of an editing complex of klenow fragment with dna. *Proc. Natl. Acad. Sci. U.S.A.* 85, 8924–8928. doi: 10.1073/pnas.85.23.8924

Gao, G. X., Orlova, M., Georgiadis, M. M., Hendrickson, W. A., and Goff, S. P. (1997). Conferring RNA polymerase activity to a DNA polymerase: a single residue in reverse transcriptase controls substrate selection. *Proc. Natl. Acad. Sci. U.S.A.* 94, 407–411. doi: 10.1073/pnas.94.2.407

Gardner, A. F., and Jack, W. E. (1999). Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res.* 27, 2545–2553. doi: 10.1093/nar/27.12.2545

Gardner, A. F., Wang, J., Wu, W., Karouby, J., Li, H., Stupi, B. P., et al. (2012). Rapid incorporation kinetics and improved fidelity of a novel class of 3′-OH unblocked reversible terminators. *Nucleic Acids Res.* 40, 7404–7415. doi: 10.1093/nar/gks330

Geyer, C. R., Battersby, T. R., and Benner, S. A. (2003). Nucleobase pairing in Watson-Crick-like genetic expanded information systems. *Structure* 11, 1485–1498. doi: 10.1016/j.str.2003.11.008

Ghadessy, F. J., Ong, J. L., and Holliger, P. (2001). Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. U.S.A.* 98, 4552–4557. doi: 10.1073/pnas.071052198

Ghadessy, F. J., Ramsay, N., Boudsocq, F., Loakes, D., Brown, A., Iwai, S., et al. (2004). Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. *Nat. Biotechnol.* 22, 755–759. doi: 10.1038/nbt974

Greagg, M. A., Fogg, A. M., Panayotou, G., Evans, S. J., Connolly, B. A., and Pearl, L. H. (1999). A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9045–9050. doi: 10.1073/pnas.96.16.9045

Hashimoto, H., Nishioka, M., Fujiwara, S., Takagi, M., Imanaka, T., Inoue, T., et al. (2001). Crystal structure of DNA polymerase from hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. *J. Mol. Biol.* 306, 469–477. doi: 10.1006/jmbi.2000.4403

Henry, A. A., and Romesberg, F. E. (2003). Beyond A, C, G and T: augmenting nature's alphabet. *Curr. Opin. Chem. Biol* 7, 727–733. doi: 10.1016/j.cbpa.2003.10.011

Hirao, I., Harada, Y., Kimoto, M., Mitsui, T., Fujiwara, T., and Yokoyama, S. (2004). A two-unnatural-base-pair system toward the expansion of the genetic code. *J. Am. Chem. Soc.* 126, 13298–13305. doi: 10.1021/ja047201d

Hirao, I., Mitsui, T., Kimoto, M., and Yokoyama, S. (2007). An efficient unnatural base pair for PCR amplification. *J. Am. Chem. Soc.* 129, 15549–15555. doi: 10.1021/ja073830m

Hollenstein, M., Hipolito, C. J., Lam, C. H., and Perrin, D. M. (2009). A DNAzyme with three protein-like functional groups: enhancing catalytic efficiency of M2+-independent RNA cleavage. *Chembiochem* 10, 1988–1992. doi: 10.1002/cbic.200900314

Hopfner, K. P., Eichinger, A., Engh, R. A., Laue, F., Ankenbauer, W., Huber, R., et al. (1999). Crystal structure of a thermostable type B DNA polymerase from *Thermococcus gorgonarius*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3600–3605. doi: 10.1073/pnas.96.7.3600

Ishikawa, M., Hirao, I., and Yokoyama, S. (2000). Synthesis of 3-(2-deoxy-β-D-ribofuranosyl)pyridin-2-one and 2-amino-6-(N,N-dimethylamino)-9-(2-deoxy-beta-D-ribofuranosyl)purine derivatives for an unnatural base pair. *Tetrahedron Lett.* 41, 3931–3934. doi: 10.1016/s0040-4039(00)00520–527

Ito, J., and Braithwaite, D. K. (1991). Compilation and alignment of dna-polymerase sequences. *Nucleic Acids Res.* 19, 4045–4057. doi: 10.1093/nar/19.15.4045

Jackel, C., Kast, P., and Hilvert, D. (2008). Protein design by directed evolution. *Annu. Rev. Biophys.* 153–173. doi: 10.1146/annurev.biophys.37.032807.125832

Joyce, C. M. (1989). How DNA travels between the separate polymerase and 3′-5′-exonuclease sites of dna-polymerase-i (klenow fragment). *J. Biol. Chem.* 264, 10858–10866.

Joyce, C. M. (1997). Choosing the right sugar: how polymerases select a nucleotide substrate. *Proc. Natl. Acad. Sci. U.S.A.* 94, 1619–1622. doi: 10.1073/pnas.94.5.1619

Joyce, C. M., and Benkovic, S. J. (2004). DNA polymerase fidelity: kinetics, structure, and checkpoints. *Biochemistry* 43, 14317–14324. doi: 10.1021/bi048422z

Joyce, C. M., and Steitz, T. A. (1994). Function and structure relationships in dna-polymerases. *Annu. Rev. Biochem.* 63, 777–822. doi: 10.1146/annurev.bi.63.070194.004021

Kim, S. W., Kim, D. U., Kim, J. K., Kang, L. W., and Cho, H. S. (2008). Crystal structure of *Pfu*, the high fidelity DNA polymerase from *Pyrococcus furiosus*. *Int. J. Biol. Macromol.* 42, 356–361. doi: 10.1016/j.ijbiomac.2008.01.010

Kim, Y., Eom, S. H., Wang, J. M., Lee, D. S., Suh, S. W., and Steitz, T. A. (1995). Crystal-structure of *Thermus-aquaticus* DNA-polymerase. *Nature* 376, 612–616. doi: 10.1038/376612a0

Kool, E. T. (2002). Replacing the nucleohases in DNA with designer molecules. *Acc. Chem. Res.* 35, 936–943. doi: 10.1021/ar000183u

Kornberg, A. (1960). Biologic synthesis of deoxyribonucleic acid. *Science* 131, 1503–1508. doi: 10.1126/science.131.3412.1503

Korolev, S., Nayal, M., Barnes, W. M., Dicera, E., and Waksman, G. (1995). Crystal-structure of the large fragment of *Thermus-aquaticus* DNA-polymerase-I At 2.5-angstrom resolution - structural basis for thermostability. *Proc. Natl. Acad. Sci. U.S.A.* 92, 9264–9268. doi: 10.1073/pnas.92.20.9264

Kuroita, T., Matsumura, H., Yokota, N., Kitabayashi, M., Hashimoto, H., Inoue, T., et al. (2005). Structural mechanism for coordination of proofreading and polymerase activities in archaeal DNA polymerases. *J. Mol. Biol.* 351, 291–298. doi: 10.1016/j.jmb.2005.06.015

Laos, R., Shaw, R., Leal, N. A., Gaucher, E., and Benner, S. (2013). Directed evolution of polymerases to accept nucleotides with nonstandard hydrogen bond patterns. *Biochemistry* 52, 5288–5294. doi: 10.1021/bi400558c

Leconte, A. M., Chen, L., and Romesberg, F. E. (2005). Polymerase evolution: efforts toward expansion of the genetic code. *J. Am. Chem. Soc.* 127, 12470–12471. doi: 10.1021/ja053322h

Leconte, A. M., Patel, M. P., Sass, L. E., Mcinerney, P., Jarosz, M., Kung, L., et al. (2010). Directed evolution of DNA polymerases for next-generation sequencing. *Angew. Chem. Int. Ed. Engl.* 49, 5921–5924. doi: 10.1002/anie.201001607

Leemhuis, H., Kelly, R. M., and Dijkhuizen, L. (2009). Directed evolution of enzymes: library screening strategies. *IUBMB Life* 61, 222–228. doi: 10.1002/iub.165

Le Grice, S. F. J., and Nowotny, M. (2014). "Reverse Transcriptases," in *Nucleic Acid Polymerases*, eds K. S. Murakami and M. A. Trakselis (New York, NY: Springer-Verlag Berlin Heidelberg), 189–214. doi: 10.1007/978-3-642-39796-7_8

Li, Y., Korolev, S., and Waksman, G. (1998). Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* 17, 7514–7525. doi: 10.1093/emboj/17.24.7514

Li, Y., Mitaxov, V., and Waksman, G. (1999). Structure-based design of *Taq* DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9491–9496. doi: 10.1073/pnas.96.17.9491

Loakes, D., Gallego, J., Pinheiro, V. B., Kool, E. T., and Holliger, P. (2009). Evolving a polymerase for hydrophobic base analogues. *J. Am. Chem. Soc.* 131, 14827–14837. doi: 10.1021/ja9039696

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* 19, 1–7. doi: 10.1093/oxfordjournals.molbev.a003973

Lutz, S. (2010). Beyond directed evolution-semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* 21, 734–743. doi: 10.1016/j.copbio.2010.08.011

Lutz, S., and Patrick, W. M. (2004). Novel methods for directed evolution of enzymes: quality, not quantity. *Curr. Opin. Biotechnol.* 15, 291–297. doi: 10.1016/j.copbio.2004.05.004

McCullum, E. O., and Chaput, J. C. (2009). Transcription of an RNA aptamer by a DNA polymerase. *Chem. Commun.* 2938–2940. doi: 10.1039/b820678c

Minakawa, N., Kojima, N., Hikishima, S., Sasaki, T., Kiyosue, A., Atsumi, N., et al. (2003). New base pairing motifs. The synthesis and thermal stability of oligodeoxynucleotides containing imidazopyridopyrimidine nucleosides with the ability to form four hydrogen bonds. *J. Am. Chem. Soc.* 125, 9970–9982. doi: 10.1021/ja0347686

Minakawa, N., Kuramoto, K., Hikishima, S., and Matsuda, A. (2006). Four hydrogen-bonding motifs in oligonucleotides. *Arkivoc* 2006, 326–337. doi: 10.3998/ark.5550190.0007.724

Moon, A. F., Garcia-Diaz, M., Batra, V. K., Beard, W. A., Bebenek, K., Kunkel, T. A., et al. (2007). The X family portrait: structural insights into biological functions of X family polymerases. *DNA Repair* 6, 1709–1725. doi: 10.1016/j.dnarep.2007.05.009

Morales, J. C., and Kool, E. T. (1999). Minor groove interactions between polymerase and DNA: more essential to replication than Watson-Crick hydrogen bonds? *J. Am. Chem. Soc.* 121, 2323–2324. doi: 10.1021/ja983502+

Nolte, F. S., Marshall, D. J., Rasberry, C., Schievelbein, S., Banks, G. G., Storch, G. A., et al. (2007). MultiCode-PLx system for multiplexed detection of seventeen respiratory viruses. *J. Clin. Microbiol.* 45, 2779–2786. doi: 10.1128/jcm.00669–667

Nour-Eldin, H. H., Hansen, B. G., Norholm, M. H. H., Jensen, J. K., and Halkier, B. A. (2006). Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res.* 34, e122. doi: 10.1093/nar/gkl635

Ohmori, H., Friedberg, E. C., Fuchs, R. P. P., Goodman, M. F., Hanaoka, F., Hinkle, D., et al. (2001). The Y-family of DNA polymerases. *Mol. Cell.* 8, 7–8. doi: 10.1016/s1097-2765(01)00278–277

Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G., and Steitz, T. A. (1985). Structure of large fragment of *Escherichia-coli* DNA-polymerase-i complexed with dTMP. *Nature* 313, 762–766. doi: 10.1038/313762a0

Ong, J. L., Loakes, D., Jaroslawski, S., Too, K., and Holliger, P. (2006). Directed evolution of DNA polymerase, RNA polymerase and reverse transcriptase activity in a single polypeptide. *J. Mol. Biol.* 361, 537–550. doi: 10.1016/j.jmb.2006.06.050

Patel, P. H., Kawate, H., Adman, E., Ashbach, M., and Loeb, L. K. (2001). A single highly mutable catalytic site amino acid is critical for DNA polymerase fidelity. *J. Biol. Chem.* 276, 5044–5051. doi: 10.1074/jbc.M008701200

Patel, P. H., and Loeb, L. A. (2000a). DNA polymerase active site is highly mutable: evolutionary consequences. *Proc. Natl. Acad. Sci. U.S.A.* 97, 5095–5100. doi: 10.1073/pnas.97.10.5095

Patel, P. H., and Loeb, L. A. (2000b). Multiple amino acid substitutions allow DNA polymerases to synthesize RNA. *J. Biol. Chem.* 275, 40266–40272. doi: 10.1074/jbc.M005757200

Piccirilli, J. A., Krauch, T., Moroney, S. E., and Benner, S. A. (1990). Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* 343, 33–37. doi: 10.1038/343033a0

Pinheiro, V. B., Taylor, A. I., Cozens, C., Abramov, M., Renders, M., Zhang, S., et al. (2012). Synthetic genetic polymers capable of heredity and evolution. *Science* 336, 341–344. doi: 10.1126/science.1217622

Pryor, J. M., Dieckman, L. M., Boehm, E. M., and Washington, M. T. (2014). "Eukaryotic Y-family polymerases: a biochemical and structural perspective," in *Nucleic Acid Polymerases*, eds K. S. Murakami and M. A. Trakselis (New York, NY: Springer-Verlag Berlin Heidelberg), 85–108.

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., Mcquillan, J. A., et al. (2012). Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* 9, 10–11. doi: 10.1038/nmeth.1814

Ramsay, N., Jemth, A. S., Brown, A., Crampton, N., Dear, P., and Holliger, P. (2010). CyDNA: synthesis and replication of highly Cy-Dye substituted DNA by an evolved polymerase. *J. Am. Chem. Soc.* 132, 5096–5104. doi: 10.1021/ja909180c

Rappaport, H. P. (1988). The 6-thioguanine 5-methyl-2-pyrimidinone base pair. *Nucleic Acids Res.* 16, 7253–7267. doi: 10.1093/nar/16.15.7253

Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008). Addressing the numbers problem in directed evolution. *Chembiochem* 9, 1797–1804. doi: 10.1002/cbic.200800298

Schmidt, M. (2010). Xenobiology: a new form of life as the ultimate biosafety tool. *Bioessays* 32, 322–331. doi: 10.1002/bies.200900147

Sefah, K., Yang, Z. Y., Bradley, K. M., Hoshika, S., Jimenez, E., Zhang, L. Q., et al. (2014). In vitro selection with artificial expanded genetic information systems. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1449–1454. doi: 10.1073/pnas.1311778111

Sherrill, C. B., Marshall, D. J., Moser, M. J., Larsen, C. A., Daude-Snow, L., and Prudent, J. R. (2004). Nucleic acid analysis using an expanded genetic alphabet to quench fluorescence. *J. Am. Chem. Soc.* 126, 4550–4556. doi: 10.1021/ja0315558

Sismour, A. M., and Benner, S. A. (2005). Synthetic biology. *Expert Opin. Biol. Ther.* 5, 1409–1414. doi: 10.1517/14712598.5.11.1409

Smith, J. M. (1970). Natural selection and concept of a protein space. *Nature* 225, 563–564. doi: 10.1038/225563a0

Staiger, N., and Marx, A. (2010). A DNA Polymerase with increased reactivity for ribonucleotides and C5-modified deoxyribonucleotides. *Chembiochem* 11, 1963–1966. doi: 10.1002/cbic.201000384

Stemmer, W. P. C. (1994). Rapid evolution of a protein in-vitro by DNA shuffling. *Nature* 370, 389–391. doi: 10.1038/370389a0

Sterner, R. (2011). Directed evolution: a powerful approach to optimising and understanding enzymes. *Chembiochem* 12, 1439–1440. doi: 10.1002/cbic.201100285

Suzuki, M., Avicola, A. K., Hood, L., and Loeb, L. A. (1997). Low fidelity mutants in the O-helix of *Thermus aquaticus* DNA polymerase I. *J. Biol. Chem.* 272, 11228–11235. doi: 10.1074/jbc.272.17.11228

Switzer, C., Moroney, S. E., and Benner, S. A. (1989). Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.* 111, 8322–8323. doi: 10.1021/ja00203a067

Tabor, S., and Richardson, C. C. (1995). A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* 92, 6339–6343. doi: 10.1073/pnas.92.14.6339

Tae, E. J. L., Wu, Y. Q., Xia, G., Schultz, P. G., and Romesberg, F. E. (2001). Efforts toward expansion of the genetic alphabet: replication of DNA with three base pairs. *J. Am. Chem. Soc.* 123, 7439–7440. doi: 10.1021/ja010731e

Tawfik, D. S., and Griffiths, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* 16, 652–656. doi: 10.1038/nbt0798-652

Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* 36, e25. doi: 10.1093/nar/gkn021

Turner, N. J. (2009). Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* 5, 568–574. doi: 10.1038/nchembio.203

Vartanian, J. P., Henry, M., and Wainhobson, S. (1996). Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Res.* 24, 2627–2631. doi: 10.1093/nar/24.14.2627

Wang, J., Sattar, A., Wang, C. C., Karam, J. D., Konigsberg, W. H., and Steitz, T. A. (1997). Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69. *Cell* 89, 1087–1099. doi: 10.1016/s0092-8674(00)80296-2

Wang, Y., Prosen, D. E., Mei, L., Sullivan, J. C., Finney, M., and Vander Horn, P. B. (2004). A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Res.* 32, 1197–1207. doi: 10.1093/nar/gkh271

Weinhold, E. G., Ellington, A., Presnell, S. R., Mcgeehan, G. M., and Benner, S. A. (1987). Evolution guidance - engineering alcohol-dehydrogenase and ribonuclease. *Protein Eng.* 1, 236–237.

Wong, T. S., Roccatano, D., and Schwaneberg, U. (2007). Challenges of the genetic code for exploring sequence space in directed protein evolution. *Biocatal. Biotransformation* 25, 229–241. doi: 10.1080/10242420701444280

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth. Internat. Congr. Genet. Ithaca N. Y.* 1, 356–366.

Wu, W., Stupi, B. P., Litosh, V. A., Mansouri, D., Farley, D., Morris, S., et al. (2007). Termination of DNA synthesis by N6-alkylated, not 3′-O-alkylated, photocleavable 2′-deoxyadenosine triphosphates. *Nucleic Acids Res.* 35, 6339–6349. doi: 10.1093/nar/gkm689

Wynne, S. A., Pinheiro, V. B., Holliger, P., and Leslie, A. G. W. (2013). Structures of an apo and a binary complex of an evolved archeal B family DNA polymerase capable of synthesising highly Cy-Dye labelled DNA. *PLoS ONE* 8:e70892. doi: 10.1371/journal.pone.0070892

Xia, G., Chen, L. J., Sera, T., Fa, M., Schultz, P. G., and Romesberg, F. E. (2002). Directed evolution of novel polymerase activities: mutation of a DNA polymerase into an efficient RNA polyrnerase. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6597–6602. doi: 10.1073/pnas.102577799

Yamtich, J., and Sweasy, J. B. (2010). DNA polymerase family X: function, structure, and cellular roles. *Biochim. Biophys. Acta.* 1804, 1136–1150. doi: 10.1016/j.bbapap.2009.07.008

Yang, G. W., Franklin, M., Li, J., Lin, T. C., and Konigsberg, W. (2002). A conserved Tyr residue is required for sugar selectivity in a pol alpha DNA polymerase. *Biochemistry* 41, 10256–10261. doi: 10.1021/bi0202171

Yuan, L., Kurek, I., English, J., and Keenan, R. (2005). Laboratory-directed protein evolution. *Microbiol. Mol. Biol. Rev.* 69, 373–392. doi: 10.1128/jmbr.69.3.393-392.2005