J|A|C|S
ARTICLES

Published on Web 09/24/2009

Evolving a Polymerase for Hydrophobic Base Analogues

David Loakes,† José Gallego,†,‡ Vitor B. Pinheiro,† Eric T. Kool,§ and Philipp Holliger*,†

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, United Kingdom, and Department of Chemistry, Stanford University, Stanford, California 94305

Received May 15, 2009; E-mail: ph1@mrc-lmb.cam.ac.uk

Abstract: Hydrophobic base analogues (HBAs) have shown great promise for the expansion of the chemical and coding potential of nucleic acids but are generally poor polymerase substrates. While extensive synthetic efforts have yielded examples of HBAs with favorable substrate properties, their discovery has remained challenging. Here we describe a complementary strategy for improving HBA substrate properties by directed evolution of a dedicated polymerase using compartmentalized self-replication (CSR) with the archetypal HBA 5-nitroindole (d5NI) and its derivative 5-nitroindole-3-carboxamide (d5NIC) as selection substrates. Starting from a repertoire of chimeric polymerases generated by molecular breeding of DNA polymerase genes from the genus *Thermus*, we isolated a polymerase (5D4) with a generically enhanced ability to utilize HBAs. The selected polymerase. 5D4 was able to form and extend d5NI and d5NIC (d5NI(C)) self-pairs as well as d5NI(C) heteropairs with all four bases with efficiencies approaching, or exceeding, those of the cognate Watson−Crick pairs, despite significant distortions caused by the intercalation of the d5NI(C) heterocycles into the opposing strand base stack, as shown by nuclear magnetic resonance spectroscopy (NMR). Unlike Taq polymerase, 5D4 was also able to extend HBA pairs such as Pyrene: ϕ (abasic site), d5NI: ϕ, and isocarbostyril (ICS): 7-azaindole (7AI), allowed bypass of a chemically diverse spectrum of HBAs, and enabled PCR amplification with primers comprising multiple d5NI(C)-substitutions, while maintaining high levels of catalytic activity and fidelity. The selected polymerase 5D4 promises to expand the range of nucleobase analogues amenable to replication and should find numerous applications, including the synthesis and replication of nucleic acid polymers with expanded chemical and functional diversity.

DNA has unique properties beyond its ability to encode genetic information, which make it an attractive supramolecular scaffold for chemistry, biotechnology and nanotechnology.[1] Despite its polyanionic backbone, it can fold into compact molecular structures forming specific receptors (aptamers) and catalysts, it can be assembled into complex nanostructures according to the well-understood rules of Watson−Crick base-pairing[2] and polymer strands of precisely defined length and sequence can be synthesized, replicated and evolved using DNA polymerases. However, the physicochemical properties of the canonical four bases span only a narrow range. Expanding the chemistry of nucleic acid polymers amenable to synthesis, replication and evolution would greatly enhance their phenotypic diversity and widen their biotechnological and clinical potential.

Hydrophobic base analogues (HBAs) potentially could provide a variety of attributes not present in the canonical bases including photoactivated chemistry, fluorescence and the potential to form novel nucleic acid structures through noncanonical stacking and hydrophobic interactions. HBAs have already found a wide range of applications in nucleic acid manipulation and hybridization and as steric or fluorescent probes of enzyme dynamics and function.[1]

In the context of nucleic acid replication, HBAs were originally studied as universal base analogues[5−7] but have been found to have a number of other intriguing properties. For example, hydrophobic isosteres of the natural bases were found to display specific pairing with the natural bases clarifying the importance of steric complementarity for replication fidelity.[1] Other HBAs were found to form specific self- and heteropairs with the potential to form an orthogonal third base-pair. For example, Romesberg and colleagues have systematically explored a large range of chemical space including substituted phenyl,[3,4] pyridyl,[5,6] and isocarbostyril,[7−9] as well as pyridones,[10] azaindoles[11] and other heterocycles[12] probing the requirements for polymerase recognition and for formation of

† MRC Laboratory of Molecular Biology.

‡ Present address: Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler 16, 46012 Valencia, Spain.

§ Stanford University.

[1] Kool, E. T. *Acc. Chem. Res.* **2002**, *35*, 936–943.

[2] Seeman, N. C. *Trends Biochem. Sci.* **2005**, *30*, 119–125.

[3] Hwang, G. T.; Romesberg, F. E. *Nucleic Acids Res.* **2006**, *34*, 2037–2045.

[4] Matsuda, S.; Henry, A. A.; Romesberg, F. E. *J. Am. Chem. Soc.* **2006**, *128*, 6369–6375.

[5] Kim, Y.; Leconte, A. M.; Hari, Y.; Romesberg, F. E. *Angew. Chem., Int. Ed.* **2006**, *45*, 7809–7812.

[6] Hari, Y.; Hwang, G. T.; Leconte, A. M.; Joubert, N.; Hocek, M.; Romesberg, F. E. *ChemBioChem* **2008**, *9*, 2796–2799.

[7] Berger, M.; Wu, Y.; Ogawa, A. K.; McMinn, D. L.; Schultz, P. G.; Romesberg, F. E. *Nucleic Acids Res.* **2000**, *28*, 2911–2914.

[8] Matsuda, S.; Romesberg, F. E. *J. Am. Chem. Soc.* **2004**, *126*, 14419–14427.

[9] Leconte, A. M.; Hwang, G. T.; Matsuda, S.; Capek, P.; Hari, Y.; Romesberg, F. E. *J. Am. Chem. Soc.* **2008**, *130*, 2336–2343.

[10] Leconte, A. M.; Matsuda, S.; Hwang, G. T.; Romesberg, F. E. *Angew. Chem., Int. Ed.* **2006**, *45*, 4326–4329.

10.1021/ja9039696 CCC: $40.75 © 2009 American Chemical Society

specific HBA self- and heteropairs. Hirao and Yokoyama have also designed a number of HBA base pairing systems based on shape complementarity and steric exclusion and some have been shown to be sufficiently orthogonal to specifically direct the incorporation of fluorophores,[13,14] biotin,[15] and iodine useful for photo-cross-linking to proteins[16] into RNA transcripts and to direct the specific incorporation of the unnatural amino acid 3-chlorotyrosine in a cell-free translation system.[17,18] Finally, large HBAs such as pyrene[19] or other derivatives of indole[20] have been found to be able to "detect" DNA damage by forming highly specific "base-pairs" with abasic sites. HBAs have thus shown clear potential to expand the chemical, functional and coding capabilities of nucleic acids for therapeutic, diagnostic or nanotechnological applications. However, their general application has been restricted by generally poor replication by natural polymerases.

Extensive synthetic efforts have been made to improve the properties of HBAs as polymerase substrates (see above) by optimizing steric fit,[1,21] inclusion of minor groove H-bond acceptors[22,23] and systematic derivatization with heteroatoms and alkyl substituents.[24] While significant progress has been made,[21] the identification of HBAs compatible with efficient enzymatic replication has remained challenging. An alternative strategy would be the engineering of the polymerase active site for improved HBA replication. Polymerases have been engineered by design,[25,26] screening[27,28] and selection.[29−31] These studies have uncovered significant plasticity in the polymerase active site for the acceptance of noncognate chemistries.[32−34] Indeed, Romesberg et al. have successfully applied their phage-based selection system to the evolution of a variant of the Stoffel fragment of Taq polymerase with 30-fold improved extension of the self-pair of propynylisocarbostyril (PICS).[34]

We have developed an alternative strategy for the evolution of polymerases, called "compartmentalized self-replication" (CSR).[31] CSR is based on a simple feedback loop, in which a polymerase replicates only its own encoding gene with compartmentalization into the aqueous compartments of a water-in-oil emulsion[35] serving to isolate individual self-replication reactions from each other. Thus, each polymerase replicates only its own encoding gene to the exclusion of those in other compartments (i.e., self-replicates). In such a system adaptive gains directly (and proportionally) translate into genetic amplification of the encoding gene.

Here we describe the application of CSR to the directed evolution of polymerases for HBA replication. Using flanking primers modified with the archetypal HBA 5-nitroindole (d5NI)[36] and its 3-carboxamide derivative (d5NIC), we performed CSR selections and isolated a polymerase (5D4) with a generically improved ability to incorporate, extend and bypass a variety of HBAs. These include a wide variety of noncognate substrates including HBA self-pairs as well as heteropairs with natural bases, abasic sites or other HBAs. Remarkably, the selected polymerase 5D4 is able to process these analogues regardless of a lack of minor groove H-bonding and despite significant distortions to the primer-template duplex structure caused by intercalation into the opposing strand base stack, as shown here for d5NI and d5NIC.

Results

1. Substrate Design and Selection. As the HBA substrate for polymerase evolution we chose the universal base analogue 5-nitroindole (d5NI). d5NI has found many uses in hybridization applications.[37] However, typically for HBAs, d5NI lacks H-bond donors or acceptors and is a very poor substrate for enzymatic replication.[38] Indeed, initial selection experiments were unsuccessful (not shown). In order to enable directed evolution, we synthesized several d5NI derivatives (not shown) to improve substrate characteristics. Among these, 5-nitroindole-3-carboxamide (d5NIC) (Figure 1a) showed promise: while virtually indistinguishable from d5NI in its effects on oligonucleotide melting temperature,[39] d5NIC proved a superior polymerase substrate, allowing some bypass, where d5NI stalled synthesis by Taq polymerase (Figure 1b) and allowing selections to proceed (see below).

We initiated polymerase selection by CSR[31] starting from the previously described polymerase library (3T)[40] prepared by molecular breeding of the *pol*A genes from three members of

(11) Tae, E. L.; Wu, Y.; Xia, G.; Schultz, P. G.; Romesberg, F. E. *J. Am. Chem. Soc.* **2001**, *123*, 7439–7440.
(12) Leconte, A. M.; Matsuda, S.; Romesberg, F. E. *J. Am. Chem. Soc.* **2006**, *128*, 6780–6781.
(13) Kawai, R.; Kimoto, M.; Ikeda, S.; Mitsui, T.; Endo, M.; Yokoyama, S.; Hirao, I. *J. Am. Chem. Soc.* **2005**, *127*, 17286–17295.
(14) Kimoto, M.; Mitsui, T.; Harada, Y.; Sato, A.; Yokoyama, S.; Hirao, I. *Nucleic Acids Res.* **2007**, *35*, 5360–5369.
(15) Moriyama, K.; Kimoto, M.; Mitsui, T.; Yokoyama, S.; Hirao, I. *Nucleic Acids Res.* **2005**, *33*, e129.
(16) Kimoto, M.; Endo, M.; Mitsui, T.; Okuni, T.; Hirao, I.; Yokoyama, S. *Chem. Biol.* **2004**, *11*, 47–55.
(17) Hirao, I.; Ohtsuki, T.; Fujiwara, T.; Mitsui, T.; Yokogawa, T.; Okuni, T.; Nakayama, H.; Takio, K.; Yabuki, T.; Kigawa, T.; Kodama, K.; Yokogawa, T.; Nishikawa, K.; Yokoyama, S. *Nat. Biotechnol.* **2002**, *20*, 177–182.
(18) Hirao, I.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Kawai, R.; Sato, A.; Harada, Y.; Yokoyama, S. *Nat. Methods* **2006**, *3*, 729.
(19) Matray, T. J.; Kool, E. T. *Nature* **1999**, *399*, 704–708.
(20) Zhang, X.; Donnelly, A.; Lee, I.; Berdis, A. J. *Biochemistry* **2006**, *45*, 13293–13303.
(21) Seo, Y. J.; Hwang, G. T.; Ordoukhanian, P.; Romesberg, F. E. *J. Am. Chem. Soc.* **2009**, *131*, 3246–3252.
(22) Morales, J. C.; Kool, E. T. *Biochemistry* **2000**, *39*, 12979–12988.
(23) Matsuda, S.; Leconte, A. M.; Romesberg, F. E. *J. Am. Chem. Soc.* **2007**, *129*, 5551–5557.
(24) Leconte, A. M.; Hwang, G. T.; Matsuda, S.; Capek, P.; Hari, Y.; Romesberg, F. E. *J. Am. Chem. Soc.* **2008**, *130*, 2336–2343.
(25) Gardner, A. F.; Jack, W. E. *Nucleic Acids Res.* **1999**, *27*, 2545–2553.
(26) Li, Y.; Mitaxov, V.; Waksman, G. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9491–9496.
(27) Suzuki, M.; Baskin, D.; Hood, L.; Loeb, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 9670–9675.
(28) Summerer, D.; Rudinger, N. Z.; Detmer, I.; Marx, A. *Angew. Chem., Int. Ed.* **2005**, *44*, 4712–4715.
(29) Jestin, J. L.; Kristensen, P.; Winter, G. *Angew. Chem., Int. Ed.* **1999**, *38*, 1124–1127.
(30) Xia, G.; Chen, L.; Sera, T.; Fa, M.; Schultz, P. G.; Romesberg, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6597–6602.
(31) Ghadessy, F. J.; Ong, J. L.; Holliger, P. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4552–4557.
(32) Ghadessy, F. J.; Ramsay, N.; Boudsocq, F.; Loakes, D.; Brown, A.; Iwai, S.; Vaisman, A.; Woodgate, R.; Holliger, P. *Nat. Biotechnol.* **2004**, *22*, 755–759.
(33) Fa, M.; Radeghieri, A.; Henry, A. A.; Romesberg, F. E. *J. Am. Chem. Soc.* **2004**, *126*, 1748–1754.
(34) Leconte, A. M.; Chen, L.; Romesberg, F. E. *J. Am. Chem. Soc.* **2005**, *127*, 12470–12471.
(35) Tawfik, D. S.; Griffiths, A. D. *Nat. Biotechnol.* **1998**, *16*, 652–656.
(36) Loakes, D.; Brown, D. M. *Nucleic Acids Res.* **1994**, *22*, 4039–4043.
(37) Loakes, D. *Nucleic Acids Res.* **2001**, *29*, 2437–2447.
(38) Smith, C. L.; Simmonds, A. C.; Felix, I. R.; Hamilton, A. L.; Kumar, S.; Nampalli, S.; Loakes, D.; Hill, F.; Brown, D. M. *Nucleosides Nucleotides* **1998**, *17*, 541–554.
(39) Too, K.; Brown, D. M.; Holliger, P.; Loakes, D. *Collect. Czech. Chem. Commun.* **2006**, *71*, 899–911.
(40) d'Abbadie, M.; Hofreiter, M.; Vaisman, A.; Loakes, D.; Gasparutto, D.; Cadet, J.; Woodgate, R.; Paabo, S.; Holliger, P. *Nat. Biotechnol.* **2007**, *25*, 939–943.
(41) Eom, S. H.; Wang, J.; Steitz, T. A. *Nature* **1996**, *382*, 278–281.

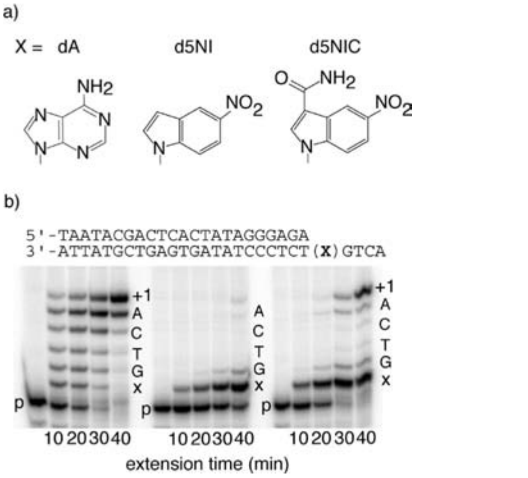

Figure 1. (a) Chemical structures of the base moieties of dA and the hydrophobic base analogues 5-nitroindole (d5NI) and 5-nitroindole-3-carboxamide (d5NIC) (b) Bypass of dA, d5NI and d5NIC by wild-type Taq polymerase. Polymerase activity was assayed for its extension ability on (from left to right) an unmodified DNA template (x = dA), the same template with d5NI and d5NIC respectively inserted at the +1 position (x = d5NI; x = d5NIC). While d5NI stalls Taq almost completely (with <10% full length extension product), there is up to 50% bypass of d5NIC.

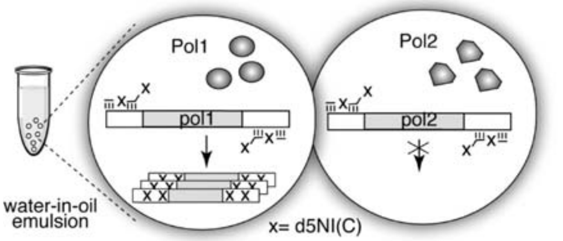

Figure 2. Principle of the CSR selection strategy with d5NI/d5NIC modified primers. CSR is based on a simple feedback loop, in which a polymerase replicates only its own encoding gene. Compartmentalization in the aqueous compartments of a water-in-oil emulsion[35] serves to isolate individual self-replication reactions from each other. Two independent emulsion compartments are shown. Polymerases (such as Pol1 (left compartment)) that are capable of utilizing flanking primers modified by d5NIC (x = d5NIC) are able to replicate their one encoding gene (*pol*1) and produce "offspring", that is, increase their copy number in the postselection population, while polymerases like Pol2 (right compartment) that are unable to utilize such primers disappear from the gene pool.

the genus *Thermus* (*Taq* (*T. aquaticus*), *Tth (T. thermophilus*), *Tfl* (*T. flavus*)).[40] To enhance and focus polymerase activity to the chosen substrates (HBAs), we modified CSR selection to include flanking primers comprising HBAs (d5NIC) both at their 3′-ends and (for later rounds of selection) internally (Figure 2), thus requiring not only HBA extension but also bypass of template HBAs for efficient self-replication.

By round two, the selected polymerase population had adapted sufficiently to d5NIC that selection with 3′-d5NI primers, too, yielded a positive selection signal (not shown). For rounds three to five we therefore carried out selections using both 3′-d5NI- respectively d5NIC-modified primers and further increased selection stringency by using primers bearing internal as well as 3′-d5NI(C) substitutions. After five rounds, the selected population comprised a diverse set of chimeric polymerases. Many of these had a strikingly improved ability to utilize d5NIC and to a lesser extent d5NI. In particular, polymerase 5D4 was able to efficiently bypass template d5NI and d5NIC (Figure 3a) as well as perform PCR with d5NIC- (and to a lesser extent d5NI)-modified primers (Figure 3b).

Most of the selected polymerases are Tth/Taq chimeras and share an arrangement of gene segments, whereby the N-terminal region (comprising part of the 5′−3′ exonuclease domain) derived from Tth, whereas the protein core derives mainly from Taq as observed previously[40] (Supporting Information Figure 1). We chose a round five polymerase 5D4, for detailed investigation. 5D4 is a Tth/Taq chimera with 14 additional mutations from the Taq/Tth consensus (V62I, Y78H, T88S, P114Q, P264S, E303V, G389V, E424G, E432G, E602G, A608V, I614M, M761T, M775T) some of which are shared among other selected polymerases (Figure 4, Supporting Information Table 1).

2. Polymerase Kinetics and Specificity. We determined single nucleotide incorporation kinetics (k_{cat}/K_m) and relative efficiencies $f_{5D4/Taq}$ for wtTaq and 5D4 using a gel-based assay[42] (Figure 5). While Taq and 5D4 displayed very similar kinetic efficiencies for the incorporation of dTTP opposite dA or dGTP opposite dC ($f_{5D4/Taq}$ = 2.6 and 1.4 respectively), 5D4 showed up to 680-fold improved incorporation kinetics of d5NICTP opposite the natural bases or incorporation of dNTPs opposite template d5NIC, approaching, or in some cases exceeding, the efficiencies of formation of the canonical base-pairs. Formation of the $[d5NI]_2$ or $[d5NIC]_2$ self-pairs by 5D4 was improved by 35 and 310-fold respectively, the latter exceeding the efficiency of formation of a dT:dA pair by the same polymerase. Indeed, all d5NI:dN/d5NIC:dN heteropairs and the $[d5NI(C)]_2$ self-pairs were formed with similar efficiencies to the formation of the canonical base-pairs (Figure 5). However, formation of the reverse dN:d5NI/dN:d5NIC heteropairs (incorporation of dNTPs opposite template d5NI/d5NIC) only reached maximal 5% of the efficiency of the formation of canonical base-pairs, despite an up to 110-fold improvement by 5D4 compared to wtTaq. Contributions to the improved catalytic efficiency of 5D4 arise from changes in both k_{cat} and K_m, with reductions in K_m for binding d5NITPs/d5NICTPs particularly striking (Supporting Information Table 2).

Once formed, basepairs involving HBAs need to be extended. While Taq could not extend either of the d5NI(C) self-pairs nor any of the d5NI(C):dN heteropairs to any detectable extent, 5D4 could efficiently extend both the d5NIC self-pair (Figure 6a) as well as the d5NI(C):dN heteropairs (Figure 6b). Extension of the d5NI self-pair, however, was very inefficient even for 5D4 (Supporting Information Figure 2a). The d5NI(C) self-pairs as well as d5NI(C):purine heteropairs are reminiscent of 3′ transversion mismatches, which are very poorly extended by Taq.[43] Indeed, we found that 5D4 could also efficiently extend a A•G transversion mismatch (Supporting Information Figure 2b).

Extension of d5NI(C):dN heteropairs by 5D4 depended not only on the identity of the template strand base (dN) ("paired" with d5NI(C)) but also (to a lesser extent) on its 5′-neighbor (dN+1). To comprehensively determine extension efficiencies, we carried out extension reactions using 3′-d5NI(C) primers in all 16 possible sequence contexts. We found that d5NI is extended best when paired with dC and to a lesser extent dA, while d5NIC extends best when paired with dT followed by dC (Figure 6b). Little extension was observed, when either analogue was paired with dG. Less striking preferences were observed for dN+1, with d5NI generally preferring dT and

[42] Creighton, S.; Bloom, L. B.; Goodman, M. F. *Methods Enzymol.* **1995**, *262*, 232–256.

[43] Huang, M.-M.; Arnheim, N.; Goodman, M. F. *Nucleic Acids Res.* **1992**, *20*, 4567–4573.

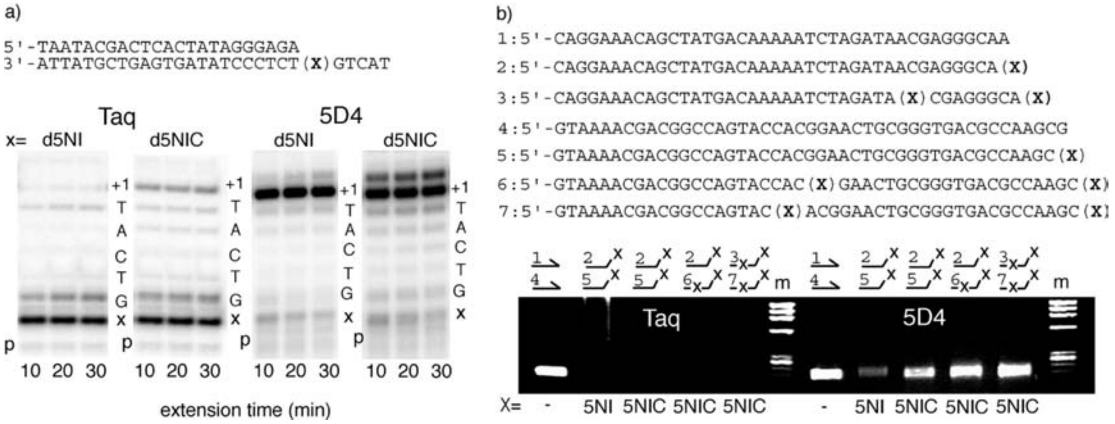

Figure 3. (a) Activity of the selected polymerase 5D4 compared to wtTaq in bypass of template d5NI and d5NIC. While Taq largely stalls, 5D4 readily bypasses both template d5NIC and d5NI. (b) PCR amplification (20× (94 °C 30 s, 50 °C 30 s, 72 °C 30 s)) using (from left to right) unmodified standard primers (1,4), or primers in which either only the 3′ terminal base is substituted by d5NI or d5NIC (2,5) or primers with 3′-d5NIC as well as internal d5NIC substitutions (3,6,7). While both polymerases show comparable activity with unmodified primers (1, 4), only 5D4 yields amplification products with d5NI(C)-modified primers (2, 5−7) including (3, 7) requiring the extension and bypass of no less than of four d5NIC substitutions for every round of replication. PCR activity with d5NI-modified primers although detectable is much poorer than with d5NIC modified primers. m, ϕX174 *Hae*III digest marker.

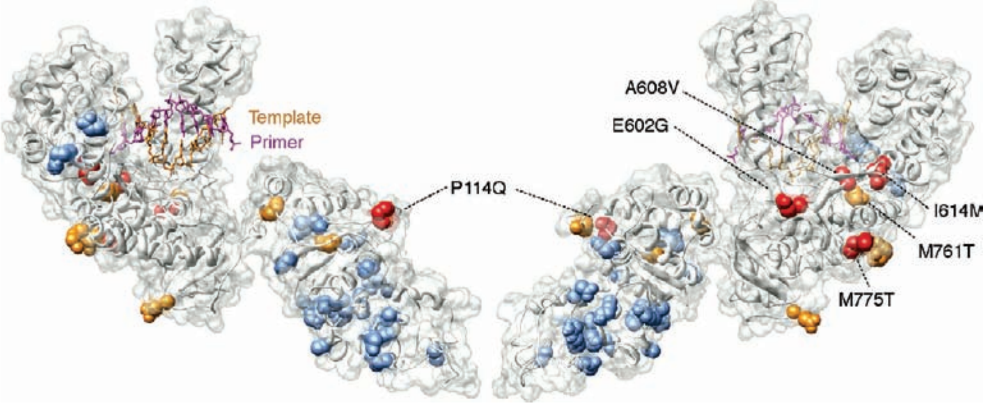

Figure 4. 5D4 polymerase structure. Mutations from the Taq consensus in the selected polymerase 5D4 are mapped on a ribbon and surface representation of Taq DNA polymerase (1TAU.pdb[41]). 5D4 mutations that derive from Tth gene segments are colored light blue. Point mutations unique to 5D4 (or shared with less than two other selected polymerases) are colored orange. 5D4 point mutations that are shared by a majority of selected polymerases (Supporting Information Table 1) are colored red.

d5NIC, dA or dC. Extension of both 3′-d5NI and 3′-d5NIC either as self-pair or as d5NI(C):dN heteropair by the wtTaq polymerase was so inefficient that no reliable kinetic constants could be deduced, while extension by 5D4 proceeded with between 1−10% the efficiency of extension of natural base-pairs pair (Supporting Information Table 3).

Together, the remarkable improvements in catalytic efficiency of the incorporation, extension and bypass of d5NI and d5NIC by 5D4 allowed PCR amplifications using primers comprising d5NI or d5NIC substitutions both at their 3′-ends as well as internally (see Figure 3b). Cloning and sequencing of PCR products allowed the determination of the coding potential of d5NI and d5NIC when replicated by 5D4 in different sequence contexts. Template d5NI predominantly directed the incorporation of dA (87%), while template d5NIC mainly templated the incorporation dT (75%) followed by dA, dG (22%) and dC (3%).

3. Generic Bypass of Hydrophobic Base Analogues. We tested the ability of 5D4 to bypass a range HBAs including 3-nitropyrrole (NP), pyrrole dicarboxamide (PDC), difluorotoluene (DFT), indole (IN), benzimidazole (BI). These include HBAs that are structurally similar to the d5NI(C) selection bait (IN, BI) as well as structurally unrelated HBAs (NP, PDC, DFT). Furthermore, we investigated the ability of 5D4 to bypass consecutive template d5NIs or d5NICs (Figure 7). While some analogues, notably PDC, NP, DFT and BI were bypassed by Taq (although poorly for BI (<10%)), bypass by 5D4 was significantly more efficient in all cases. Indeed, neither template IN nor tandem d5NI-d5NI or d5NIC-d5NIC showed detectable bypass by Taq, while all three could be bypassed by 5D4. Incorporation specificities of dNTPs opposite HBAs show distinct preferences for Taq (predominantly following the A-rule[44−46] while 5D4 displays much reduced bias (Supporting Information Figure 3).

[44] Sagher, D.; Strauss, B. *Biochemistry* **1983**, *22*, 4518–4526.
[45] Randall, S. K.; Eritja, R.; Kaplan, B. E.; Petruska, J.; Goodman, M. F. *J. Biol. Chem.* **1987**, *262*, 6864–6870.
[46] Takeshita, M.; Chang, C. N.; Johnson, F.; Will, S.; Grollman, A. P. *J. Biol. Chem.* **1987**, *262*, 10171–10179.
[47] Zahn, K. E.; Belrhali, H.; Wallace, S. S.; Doublie, S. *Biochemistry* **2007**, *46*, 10551–10561.
[48] Tae, E. L.; Wu, Y.; Xia, G.; Schultz, P. G.; Romesberg, F. E. *J. Am. Chem. Soc.* **2001**, *123*, 7439–7440.
[49] Yu, C.; Henry, A. A.; Romesberg, F. E.; Schultz, P. G. *Angew. Chem., Int. Ed.* **2002**, *41*, 3841–3844.

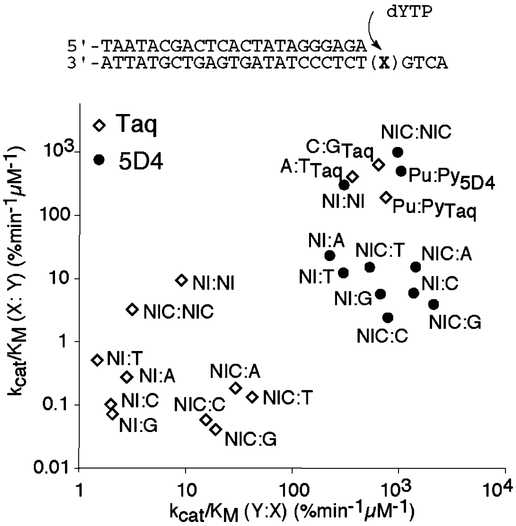

Figure 5. Steady-state kinetics $k_{\mathrm{cat}}/K_{\mathrm{M}}$ values for incorporation of dNTP and d5NI(C)TP opposite dN and d5NI(C) template bases are plotted for both Taq (◇) and 5D4 (●). Kinetic values for formation of cognate base-pairs are comparable for Taq and 5D4, while d5NI(C):X heteropairs are formed between 2−3 orders of magnitude more efficiently by 5D4 than Taq, with formation of d5NI(C) self-pairs reaching or exceeding the formation of cognate base pairs both by Taq and 5D4. Kinetic values for Pu:Py pairs for Taq and 5D4 denote the average measured for formation of a G:C and a T:A pair (Supporting Information Table 2).

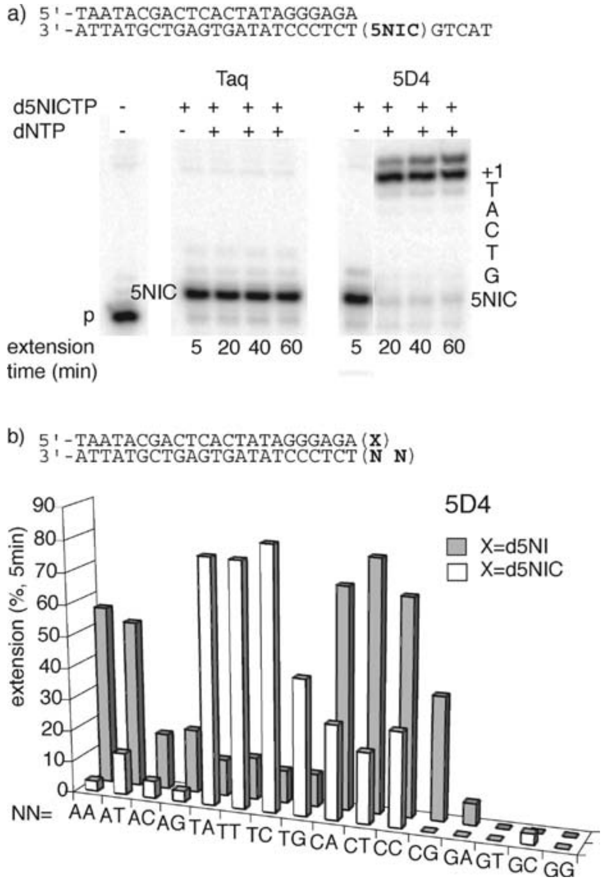

Figure 6. (a) Formation and extension of a d5NIC:d5NIC self-pair is shown for both Taq and 5D4. While d5NIC:TP is incorporated efficiently opposite template d5NIC by both polymerases, further extension of the self-pair by Taq is stalled, while 5D4 efficiently extends the self-pair. (b) Extension of 3′-d5NI(C) (% fully extended primer, 5 min) by 5D4 is plotted versus template bases N, N+1.

4. Generic Extension of Specific Base-pairs Involving HBAs. A number of HBAs can be incorporated efficiently and specifically by natural polymerases opposite their cognate partners but the nascent basepair then cannot be further extended and acts as a terminator. Cases in point are Pyrene (as its *C*-nucleoside triphosphate, dPyTP),[24,50,51] the triphosphate of d5NI (d5NITP) as well as and a number of other indole derivatives.[25,52,53] They have been shown to be incorporated with remarkable efficiency and specificity opposite a tetrahydrofuran abasic site analogue (ϕ). In both cases, incorporation is more efficient than "default" incorporation of dATP according to the A-rule,[44−46] presumably due to good steric complementarity to the missing base pair,[47] and strong π-stacking with the 5′-nucleotide. However, in both cases further extension of the dPy:ϕ or d5NI:ϕ pairs by naturally occurring polymerases is absent or very inefficient. Indeed, we find that while dPyTP and d5NITP are incorporated efficiently opposite ϕ by Taq polymerase, it is then unable to extend beyond the dPy:ϕ or d5NI:ϕ pairs (3′-base: template base) (Figure 8). In contrast, 5D4 could both form and extend the dPy:ϕ or d5NI:ϕ pairs with good efficiency. 5D4 also could bypass a template ϕ by incorporating dATP opposite the abasic site and extending the dA:ϕ pair (not shown), but, surprisingly, extension of a dPy:ϕ pair was superior than that of a dA:ϕ pair (Figure 8).

We also examined 7-azaindole (7AI) and isocarbostyril (ICS), which had been reported to form specific self- and heteropairs,[48,49] which were poorly extended by natural polymerases. Indeed, we found extension of ICS:ICS, ICS:7AI, 7AI:ICS and 7AI:7AI pairs by Taq to be close to undetectable (<3%). In contrast, 5D4 could extend all pairs to some degree, with extension improved significantly by (1 mM) Mn^{2+}. Despite the structural similarity of 7AI and 5NI, extension of 3′-7AI was generally weak and extension of 3′-ICS clearly favored. The ICS:7AI pair proved a particularly good substrate with extension reaching completion within 5 min with little or no stalled intermediates (Figure 9). It should be noted that superior HBA heteropairs have since been described since by the Romesberg lab[21] but these were not accessible to us. Our use of ICS and 7AI was to illustrate the potential of 5D4 for improving the utilization of HBAs that are inherently suboptimal polymerase substrates.

Structural Analysis of DNA Primer-Template Duplexes Containing d5NI and d5NIC. To better understand the structural framework for d5NI(C) incorporation and extension, we determined the solution structure of two primer-template duplexes (*tni* and *tnic*) with d5NI or d5NIC poised for extension at the 3′-end of the primer strand by NMR spectroscopy (Figure 10a).

The most striking deviation from canonical DNA structure is the intercalation of the d5NI and d5NIC heterocycles into the template strand base-stack: in both *tni* and *tnic*, the nitroindole rings are unpaired and stack on the neighboring C14:G5 pair, intercalating between template nucleotides A4 and G5 (Figure 10b). This conformation is clearly indicated by an extensive network of sugar-aromatic and aromatic−aromatic NOEs between d5NI(C)15 and the C14, A4 and G5 nucleotides, and by the absence of the expected A4 H2-G5 H1′ interaction (Figure 11a, Supporting Information Figure 4).

(50) Gallego, J.; Loakes, D. *Nucleic Acids Res.* **2007**, *35*, 2904–2912.

(51) Klewer, D. A.; Zhang, P.; Bergstrom, D. E.; Davisson, V. J.; LiWang, A. C. *Biochemistry* **2001**, *40*, 1518–1527.

(52) Smirnov, S.; Matray, T. J.; Kool, E. T.; de los Santos, C. *Nucleic Acids Res.* **2002**, *30*, 5561–5569.

(53) Reineks, E. Z.; Berdis, A. J. *Biochemistry* **2004**, *43*, 393–404.

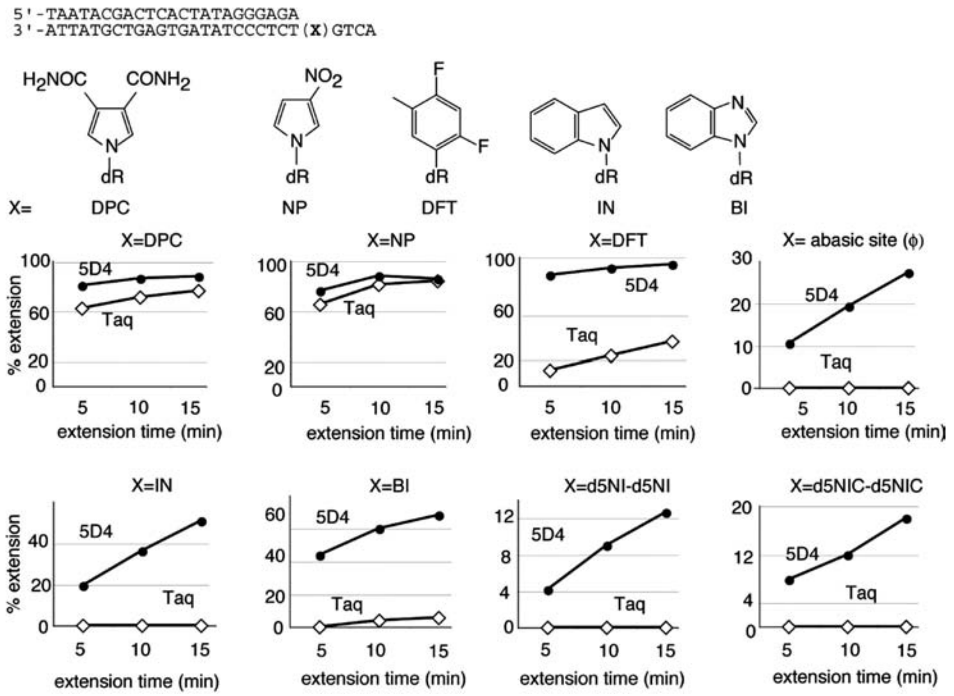

Figure 7. Bypass of hydrophobic and universal base analogues. Primer extension (% of fully extended versus input primer) on DNA templates comprising hydrophobic and universal base analogues X = DPC, NP, DFT, IN, BI at the +1 position (and the +1 and +2 position for X = d5NI-d5NI, d5NIC-d5NIC) is plotted against incubation time for activity normalized polymerase concentrations of wtTaq (◇) and 5D4 (●). Chemical structures for DPC, NP, DFT, IN, BI are shown. For the small HBAs, DPC and NP, Taq and 5D4 show similar activity, while 5D4 greatly out-performs Taq on larger HBAs. In particular 5D4 is able to bypass analogues such as IN, BI, $[d5NI(C)]_2$ as well as an abasic site that stall Taq under the experimental conditions.

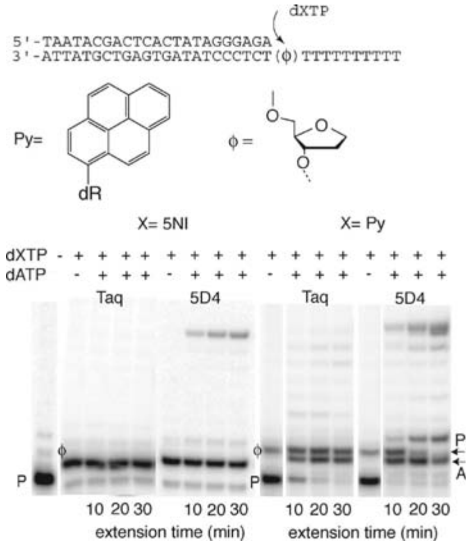
dXTP
5'-TAATACGACTCACTATAGGGAGA
3'-ATTATGCTGAGTGATATCCCTCT(ϕ)TTTTTTTTTT
Py=
dR
ϕ =
X= 5NI
X= Py
dXTP - + + + + + + + + + + + + + + + +
dATP - + + + - + + + - + + + - + + +
Taq
5D4
ϕ
P
Py:ϕ
A:ϕ
10 20 30
extension time (min)

Figure 8. Formation and extension of base-pairs between large HBAs and an abasic site. Primer extension on DNA templates comprising an abasic site (ϕ) at the +1 position by Taq and 5D4 is compared for incorporation of d5NITP (left) and dPyTP (right), followed by dATP chase. Chemical structures for the pyrene base (Py) and the tetrahydrofuran abasic site (ϕ) are shown. Incorporation of d5NITP and dPyTP is efficient for both Taq and 5D4. However, Taq is unable to extend neither the d5NI:ϕ nor the dPy:ϕ pair, while 5D4 is able to extend both. In the case of dPy, extension with dATP leads to the formation of a dA:ϕ pair on the unoccupied ϕ sites (lower band) by both Taq and 5D4. This dA:ϕ pair is also not extended by Taq, while it is extended by 5D4, although, remarkably, 5D4 preferentially extends the dPy:ϕ pair (upper band). P is the unextended primer.

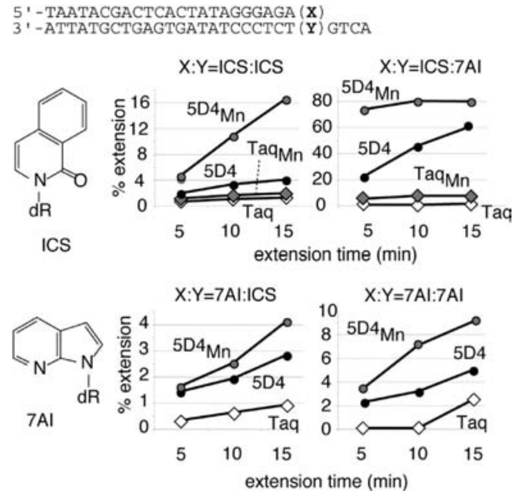

Figure 9. Extension of performed HBA base-pairs. Primer extension (% of fully extended product vs input primer) on DNA templates with hydrophobic base-pairs comprising isocarbostyril (ICS) and 7-azaindole (7AI) self- (ICS:ICS; 7AI:7AI) and heteropairs (ICS:7AI, 7AI:ICS) is plotted against incubation time for activity normalized polymerase concentrations of wtTaq (open diamonds) and 5D4 (closed circles) in the presence and absence of 1 mM Mn^{2+} (Taq_{Mn}, (gray diamonds); $5D4_{Mn(}$(gray circles)). While Taq only poorly extends base-pairs comprising 3′-ICS (ICS:ICS, ICS:7AI), 5D4 is able to extend these (especially the ICS:7AI heteropair). Extension unnatural base-pairs comprising 3′-7AI (7AI:ICS, 7AI:7AI) by 5D4 is less efficient but superior to Taq.

As observed previously,[50] both d5NI and d5NIC adopt standard *anti* conformations (Figure 11a, b), placing the H2 and H3 protons (purine numbering) in the minor groove of the helix and H6, H8 and H7 (d5NI) or the carboxamide group (d5NIC) in the major groove. The d5NIC15 carboxamide is oriented such that the amide-NH_2 group points toward the primer strand backbone and establishes a hydrogen-bonding interaction with the d5NIC phosphate (Figure 10b). This contact is supported by the observation of two sharp downfield- and upfield-shifted carboxamide proton resonances giving rise to stronger NOE interactions with d5NIC H8 relative to d5NIC H6, and to NOEs

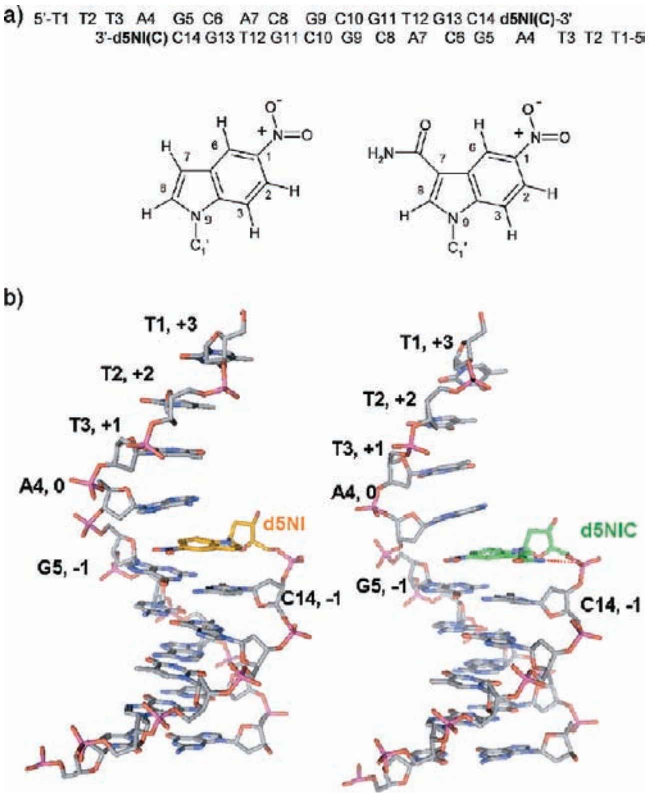
a)
5'-T1 T2 T3 A4 G5 C6 A7 C8 G9 C10 G11 T12 G13 C14 d5NI(C)-3'
3'-d5NI(C) C14 G13 T12 G11 C10 G9 C8 A7 C6 G5 A4 T3 T2 T1-5'
H
O⁻
+N=O
H 7 6 1
8 2 H
H N 9 3
C1'
H
H2N
O
b)
T1, +3
T2, +2
T3, +1
A4, 0
d5NI
G5, -1
C14, -1
d5NIC

Figure 10. NMR spectroscopy analysis of primer-template duplexes containing d5NI and d5NIC base analogues. (a) sequence of the symmetric primer-template duplexes studied by NMR spectroscopy. (b) NMR models of primer-template duplexes containing d5NI (*tni*) and d5NIC (*tnic*) at the 3′-end of the primer strand. For clarity only one-half of the symmetric duplex structures are shown. The models are oriented to illustrate the stacking of d5NI (orange) and d5NIC (green) on the −1 base-pair (G5:C14) and the intercalation of the nitroindole heterocycles between A4 and G5 of the template strand. The carboxamide group of d5NIC is oriented toward the C14-d5NIC15 backbone and establishes a hydrogen-bonding interaction with the d5NIC15 phosphate (dotted red line).

with the sugar H3′, H2′ and H2″ and base H5 and H6 protons of C14, with the downfield-shifted proton generating stronger NOEs with d5NIC H8 and C14 H3′ (Figure 11b). The interaction between the downfield-shifted carboxamide proton and the d5NIC15 phosphate covalently bonded to C14 O3′ is further supported by the adoption of an unusual C3′-*endo* conformation by the *tnic* (but not *tni*) C14 sugar, which facilitates the formation of this intrastrand contact. The absence of any unusual resonance broadening in *tni* and *tnic* contrasts with the dynamic effects previously observed in DNA duplexes containing d5NI and d5NIC in internal positions, where the unpaired nitroindole bases were found to exchange between two alternative intercalated conformations:[50] in *tni* and *tnic* no spectral broadening is observed because intercalation between A4 and G5 is the only stable stacked conformation available for the terminal nitroindole base (Figure 10b).

Discussion

We have used CSR selection[31] for the directed evolution of a polymerase (5D4) with a generically enhanced ability for the synthesis of nucleic acids comprising HBAs, including large HBA analogues that display poor geometric fit and lack minor groove H-bonding capacity, which were previously refractive to enzymatic incorporation, extension and/or bypass. Examples for such HBAs are d5NI and d5NIC, which were used here as

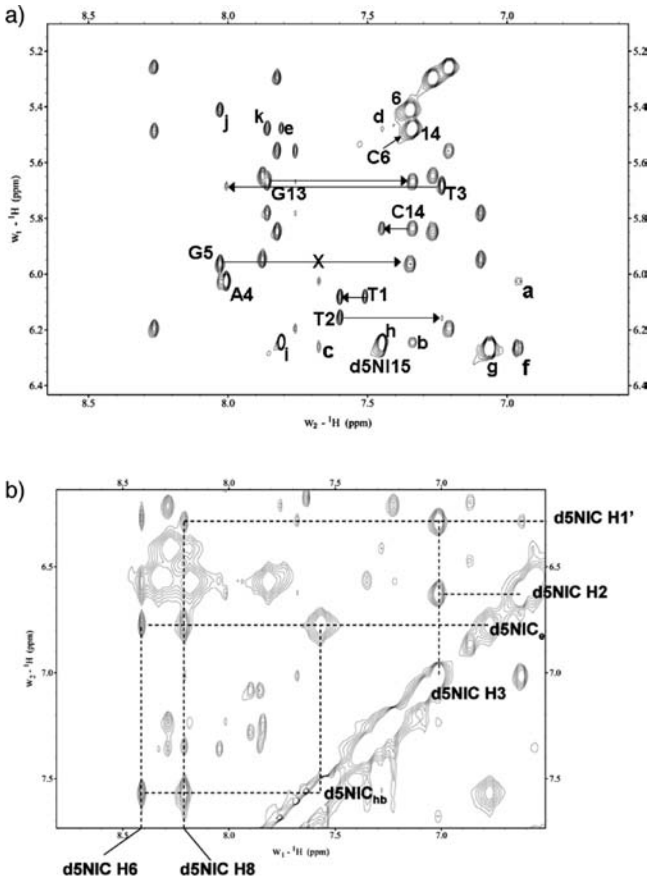

Figure 11. (a) Aromatic-H1′ region of the D_2O NOESY spectra (250 ms mixing time/25 °C) of *tni*. Intraresidue pyrimidine H6−H1′ and purine/d5NI H8−H1′ cross-peaks are labeled with residue name and number, intraresidue pyrimidine H5−H6 crosspeaks are labeled with residue number, and sequential NOE connectivities are indicated with arrows, (only T1 to C6 and G13 to d5NI15 assignments shown). Cross-peaks a−k are assigned as follows: (a) d5NI15 H2-A4 H1′; (b) C14 H6-d5NI15 H7; (c) A4 H2-d5NI15 H1′ and A4 H2-d5NI15 H7 (overlapped); (d) d5NI15 H8−C14 H5; (e) d5NI15 H6−C14 H5; (f) d5NI15 intraresidue H2−H1′ and H2−H7 (overlapped); (g) d5NI15 intraresidue H3−H1′ and H3−H7 (overlapped); (h) d5NI15 intraresidue H7−H8 (overlapped with H8−H1′); (i) d5NI15 intraresidue H7−H6; (j) G5 H8−C6 H5; (k) G13 H8−C14 H5. The absence of the sequential A4 H2-G5 H1′ NOE is marked with an "X". (b) Aromatic region of the H_2O NOESY spectra (120 ms mixing time/9 °C) of *tnic*, showing the interactions between the d5NIC H2, H3, H6, H8 and carboxamide (e and hb) protons.

the selection "bait". d5NI is an "archetypal" HBA in that it lacks any H-bonding potential and, consequently, interacts with the opposing base by stacking (rather than pairing with it), giving rise to its universal base properties in hybridization applications.[36] Typically for many larger HBAs, this stacking interaction causes conformational distortions of the primer-template duplex by intercalation into the opposing strand base-stack as we have shown here by NMR (Figure 10b). While d5NI and d5NIC display virtually identically hybridization properties, d5NIC is a superior polymerase substrate. Examination of the NMR data suggests a possible mechanism for the enhanced rate of d5NIC incorporation and extension. We find the carboxamide group of d5NIC projecting into the major groove allowing us to exclude a potential participation in minor groove H-bonding, which crucially affects polymerase extension efficiency.[22] Location in the major groove may be preferred due to its hydrophilic nature, which permits the solvation of the carboxamide group. Indeed, Klewer et al. studied the NMR structure of duplexes containing another carboxamide substituted HBA

(1,2,4-triazole-3-carboxamide) and found that the carboxamide group also resided in the major groove.[51] However, while unable to participate in minor groove interactions with the polymerase, we find the carboxamide group perfectly placed for a hydrogen bonding interaction with its own 5′-phosphate backbone group (Figure 10b). In addition, we find that 5-nitroindole-3-methylcarboxamide, an analogue with similar stacking but in which the ability of the carboxamide group to form hydrogen bonds is hindered, is as poor a polymerase substrate as d5NI (not shown). This suggests an important contribution from this hydrogen bond interaction toward d5NIC's favorable substrate properties, presumably by restriction of lateral movement and improved positioning of the 3′−OH primer terminus for catalysis. Design of proximal H-bonding groups into HBAs may be worth exploring as a general strategy to improve properties of HBAs or indeed other base analogues for enzymatic replication.

Another potential interaction is observed when d5NIC is "paired" with dT in the opposite strand (as opposed to dA as shown in Figure 10). In that case, in addition to the intrastrand hydrogen-bond to its own phosphate described here, the carboxamide group can form an out-of-plane interstrand hydrogen-bond with O4 of dT.[50] The latter may provide an explanation for the different sequence bias of d5NIC extension (Figure 6) and templating (favoring dT incorporation) compared to d5NI (favoring dA). However, this interaction does not appear to play any role during the incorporation step as neither the incorporation of d5NICTP opposite template dT nor the incorporation of dTTP opposite template d5NIC is especially favored (Figure 5).

Despite these noncanonical interstrand and intrastrand interactions by d5NIC and the distortions caused by the intercalation of both d5NI and d5NIC into the template strand base-stack, both the $[\text{d5NI(C)}]_2$ self-pairs as well some of the d5NI(C):dA, dG, dC, dT heteropairs are synthesized by 5D4 with kinetic efficiencies approaching or exceeding those of the canonical base-pairs (Figure 5). Once formed, d5NI(C):dN heteropairs and d5NIC self-pairs (but not d5NI self-pairs) are efficiently extended by 5D4, while neither is extended by Taq (Figure 6, Supporting Information Figure 2). 5D4 also greatly outperforms other polymerases on HBA pairs that do not distort DNA conformation, notably the dPy and d5NI heteropairs with an abasic site (ϕ). In these, the HBA ring occupies the space left by the missing template base and completes the opposing strand base-stack without distorting DNA conformation.[47,52] While formed efficiently and specifically by natural polymerases, they act as terminators.[19,53] 5D4, in contrast, is not only able to extend the unnatural d5NI:ϕ and dPy:ϕ "base-pairs" efficiently but even extends dPy:ϕ in preference to a "natural" dA:ϕ pair (Figure 8). Specific formation and efficient extension of dPy:ϕ and d5NI:ϕ heteropairs by 5D4 raises the potential of the synthesis of long DNA polymers with d5NI, dPy (and potentially other large HBAs) inserted at defined positions as determined by the positioning of ϕ groups in the synthetic template.

The ability of 5D4 to efficiently replicate a wide variety of HBAs allowed their potential for coding to be readily explored. We examined the coding preferences of 3-nitropyrrole (NP), pyrrole dicarboxamide (PDC), difluorotoluene (DFT), indole (IN), benzimidazole (BI), two consecutive template d5NIs or d5NICs as well as the tetrahydrofuran abasic site analogue (ϕ) with both wtTaq and 5D4. We find that, when Taq is able to bypass an HBA it predominantly follows the A-rule[44−46] due to dA's favorable stacking properties. Although generally favoring dATP and dTTP incorporation, 5D4 displays a much more even incorporation profile, approaching near universal base behavior for NP and PDC both as templating bases and as deoxynucleotide triphosphates (Supporting Information Figures 3 and 5).

5D4 is a chimeric polymerase incorporating segments from both Taq and Tth polymerases, as well as a very short segment from Tfl at the N-terminus. No crystal structures exist for either Tth or Tfl but, with on average 80% sequence homology between the three polymerases, the available structures of Taq polymerase provide a close structural analogue for those regions deriving from Tth and Tfl. Due to its chimeric nature, 5D4 differs by a total of 41 mutation from the Taq consensus. The bulk of these mutations is concentrated in the 5′−3′ exonuclease domain, which largely derives from Tth, while the main polymerase domain of 5D4 largely derives from Taq except for two short Tth segments around residues 710−730 and at the very C-terminus (Supporting Information Figure 1). In addition to the mutations deriving from Tth, 5D4 comprises 14 point mutations not present in the parental genes. Some of these are unique to 5D4, while others are shared with a group of polymerases isolated from CSR rounds 4 (4C11) and 5 (5B1, 5B4, 5D3) (Supporting Information Figure 1), which display a very similar (if slightly weaker) phenotype than 5D4. This identifies E602G, A608V, I614M, M762T and M775T as mutations within the main polymerase domain that are likely to be associated with the phenotype as they are present in all or most of the polymerases from this group (Supporting Information Table 1). However, from a simple inspection of Taq polymerase structure it is far from clear how these mutations contribute to the phenotype as, with the exception of I614M, they are distant from the active site. We attempted to rationalize the relative contributions of these mutations (as well as others) toward the 5D4 phenotype using computational analysis as well as reverse genetics.

By their iterative nature, directed evolution experiments frequently yield mutations that incrementally contribute to the new phenotype. Such mutations are often located distal to the active site as direct perturbation of the active site would be most likely to cause substantial losses in catalytic activity and be selected against. Various strands of evidence suggest that such mutations can affect the functional properties of active sites by correlated motion propagated through networks of amino acids,[54] which connect distal regions of the protein structure with each other. An effective approach to identify such networks is based on the assumption that covarying residues in protein families reveal functionally interacting amino acids independent of their location in the three-dimensional protein structure[55,56] and may thus be a useful tool to rationalize the outcome of evolution experiments as suggested by Lockless and Muir.[57] We have applied this approach, utilizing Statistical Coupling Analysis (SCA) to the main polymerase domain using an alignment of 994 members of the *pol*A family of bacterial DNA polymerases. Hierarchical clustering analysis using a high stringency cutoff to map only the strongest correlations, identified 40 residues (10% of the aligned sequence) that together with the conserved polymerase core (>97% conservation across all sequences), form

[54] Goodey, N. M.; Benkovic, S. J. *Nat. Chem. Biol.* **2008**, *4*, 474–482.

[55] Lockless, S. W.; Ranganathan, R. *Science* **1999**, *286*, 295–299.

[56] Suel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. *Nat. Struct. Biol.* **2003**, *10*, 59–69.

[57] Lockless, S. W.; Muir, T. W. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10999–11004.

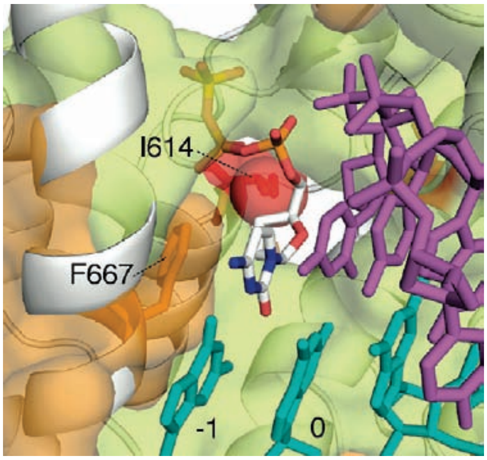

Figure 12. Structure of the polymerase active site: Structure of the active site of the closed ternary complex of Taq polymerase (3KTQ[64]) with the mutated residue I614 shown in red. Conserved residues (>97% identity) are shown as the green surface together with the SCA network, shown as the orange surface comprising F667. The template strand is shown in teal and nascent primer strand in purple. Both I614 and F667 make close contacts with the incoming dNTP, shown in elemental colors.

a network of spatially contiguous amino acids and display multiple contact points to the primer-template duplex (Supporting Information Figure 6). SCA analysis revealed M761T and M775T (as well as E734N) as part of the network suggesting a conduit by which they could modify polymerase function. The uncovered network connects these mutations to F667 within the polymerase active site in direct contact with the incoming deoxynucleotide triphosphate (Figure 12). Numerous studies have identified F667 (or its equivalent residue F762 in *E. coli* DNA polymerase I) as a key factor in polymerase substrate selection. Mutation of F667 has been shown to have dramatic effects e.g. in the discrimination against ddNTPs as well as on nucleotide incorporation fidelity.[58,59] Its function has been proposed to be to constrain the incoming dNTP molecule as part of geometric substrate selection and position it correctly for attack by the primer 3′-OH.

It is important to note that mutations may be selected for entirely different reasons than a direct contribution to HBA utilization. During directed evolution experiments, other traits are also under adaptive pressure. These include expression, folding and in particular protein stability. As the majority of mutations are likely to be destabilizing, there is a limit to the number of mutations that can be sustained before a critical stability threshold is reached and the protein is no longer functional under the selection conditions.[60] This is of especially acute importance in selection regimes such as CSR, where the protein is subjected to high temperature thermocycling conditions. We had previously observed that chimeric polymerases comprising the Tth 5′−3′ exonuclease domain and the Taq polymerase domain were substantially more thermostable than Taq polymerase.[40] The chimeric nature of 5D4 and the other selected polymerases, comprising a large segments of the Tth 5′−3′ exonuclease domain may therefore have been selected for due to its stabilizing effect on overall polymerase structure, thereby promoting evolvability through increased tolerance of destabilizing mutations. Indeed, using FoldX analysis to predict the change in free energy of folding ($\Delta\Delta G$) upon introduction of 5D4 specific mutations into the Taq framework, we found that most mutations in the main polymerase domain are destabilizing (Supporting Information Figure 7, Table 5). However, FoldX[61] analysis also identified A608V and surprisingly I614M as key stabilizing mutations shared by all selected polymerases. A608V was previously observed in a mutant Taq polymerase (T8) selected for increased thermostability.[31] We therefore conclude that the bulk of mutations in the 5′-3′exonuclease domain together with A608V are likely to have been selected to increase overall polymerase stability and offset the destabilizing influence of other mutations of adaptive value.

We reverted the E602G, I614M, M762T and M775T mutations in the main polymerase domain that were conserved among all the polymerases displaying the “HBA phenotype” back to wild-type (G602E, M614I, T762M, T775M) and analyzed their properties using the d5NIC PCR assay (Figure 3b) as it simultaneously determines d5NIC extension and bypass ability and most closely resembles the CSR process by which the mutations were selected. Only one backmutation, 5D4: M614I, displayed a significant reversion phenotype, in that its activity in d5NIC PCR was markedly reduced (Supporting Information Figure 8). All the other back mutations showed only marginal reductions in d5NIC PCR activity and may therefore contribute only incrementally toward the phenotype or through interaction with I614M. I614 is located in the A-motif within the polymerase active site and is directly involved in binding the incoming dNTP substrate. The change from Ile to Met, results in decreased steric constraints within the active site by removal of a CH_3 group projecting into the active site and toward the incoming dNTP (Figure 12). Indeed, mutation of I614 (I614K,[62] I614M[63] or I614T[30]) has been found to decrease discrimination against noncognate substrates such as NTPs within the polymerase active site either alone or in conjunction with mutation of the juxtaposed E615 steric gate residue. Within the same group of selected polymerases the proximal residue E602 has also been found to be mutated (E602V[63]) in conjunction with I614M.

Taken together, these findings clearly implicate I614 as a critical residue in the steric control of substrate selection and suggest that the 5D4 phenotype may arise to a substantial extent from a simple relaxation of steric control within the active site through the I614M mutation as well as through the propagation of the effects of the M761T and M775T (and E734N) mutations through the SCA network to the polymerase active site.

One prediction arising from such a model would be that Taq and 5D4 should perform approximately equally well on small HBA substrates, while 5D4 should outperform Taq on large HBAs. This is exactly what we observe. For example, while bypass of small template HBAs like DPC and NP proceeds with comparable efficiency for Taq and 5D4 (Figure 7), only 5D4 is able to bypass larger HBAs such as IN, BI and d5NI(C). Likewise, incorporation of small HBA triphosphates like NP-TP and PMC-TP proceeds with comparable efficiency for Taq and 5D4 (Supporting Information Figure 5), while incorporation

[58] Astatke, M.; Grindley, N. D.; Joyce, C. M. *J. Mol. Biol.* **1998**, *278*, 147–165.

[59] Suzuki, M.; Yoshida, S.; Adman, E. T.; Blank, A.; Loeb, L. A. *J. Biol. Chem.* **2000**, *275*, 32728–32735.

[60] Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5869–5874.

[61] Guerois, R.; Nielsen, J. E.; Serrano, L. *J. Mol. Biol.* **2002**, *320*, 369–387.

[62] Patel, P. H.; Loeb, L. A. *J. Biol. Chem.* **2000**, *275*, 40266–40272.

[63] Ong, J. L.; Loakes, D.; Jaroslawski, S.; Too, K.; Holliger, P. *J. Mol. Biol.* **2006**, *361*, 537–550.

[64] Li, Y.; Korolev, S.; Waksman, G. *EMBO J.* **1998**, *17*, 7514–7525.

of larger HBA-TPs like d5NI(C)-TP by 5D4 is up to 300-fold more efficient than Taq (Figure 5). Incorporation of dyATP,[65] a large expanded dA analogue with Watson−Crick H-bonding ability but poor geometric fit and extension of a distorting A•G transversion mismatch were also enhanced by 5D4 (Supporting Information Figures 2b and 9). The relaxation of geometric substrate selection also permits increased tolerance for noncognate template strand conformations. As we have shown here by NMR (Figure 10, 11), 3′-d5NI(C) causes distortion of the template strand conformation by intercalation. Nevertheless, 3′-d5NI(C):dN heteropairs (and d5NIC:d5NIC self-pairs) are efficiently extended by 5D4 while they stall extension by Taq polymerase (Figure 6).

Relaxing steric control in the polymerase active site might also be expected to give rise to low fidelity, poor catalytic efficiency and reduced processivity. However, we find dNTP incorporation and extension kinetics (Figure 5) as well as efficiency in standard PCR to be comparable for Taq and 5D4 (Figure 3b). Similarly, we find that, although the overall rate of nucleotide misincorporation (3.1×10^{-4}) by 5D4 is increased ca. 5-fold compared to wtTaq polymerase (M. Arana, PH, T. Kunkel, manuscript in preparation), it is comparable to other polymerases (such as Klenow exo^{-}[66]) widely used in molecular biology applications.

In conclusion, CSR selection using oligonucleotide primers comprising the HBA d5NI and its carboxamide derivative d5NIC as substrates has yielded 5D4, a polymerase with a generic ability to synthesize nucleic acids comprising HBAs while maintaining robust catalytic activity and fidelity. Particularly striking is the capacity of 5D4 to form and extend a diverse collection of unnatural base pairs involving HBAs. The ability of 5D4 to efficiently process large analogues that lack minor groove H-bonding and distort cognate DNA geometry should relax HBA design constraints and expedite the synthesis of DNA fragments comprising diverse HBAs. The properties of 5D4 bode well for its application in unlocking the coding potential of HBAs and other unnatural nucleotide analogues, previously incompatible with enzymatic replication.

Materials and Methods

Nucleotides and Oligonucleotides. 5-Nitroindole and difluorotoluene phosphoramidites were supplied by Glen Research. 5-Nitroindole-3-carboxamide phosphoramidite,[39] 1-(2-deoxy-β-D-ribofuranosyl)-5-nitroindole 5′-triphosphate,[38] pyrrole dicarboxamide,[67] indole, benzimidazole,[68] pyrene[19] and dyATP[65] were prepared as previously described. 5-Nitroindole-3-carboxamide attached to controlled pore glass support was prepared according to the method of Pon.[69]

Synthesis of 1-(2-Deoxy-β-D-ribofuranosyl)-5-nitroindole-3-carboxamide 5′-Triphosphate. To an ice-cold solution of methyl-1-(2-deoxy-β-D-ribofuranosyl)-5-nitroindole-3-carboxylate[39] (100 mg, 0.3 mmol) and proton sponge (96 mg, 0.45 mmol) in trimethyl phosphate (3 cm^3) was added phosphoryl chloride (35 μL, 0.38 mmol) and the solution stirred at 0 °C for 5 h. To this was added simultaneously tributylamine (0.5 cm^3) and tetrabutylammonium pyrophosphate solution (0.5 M in DMF, 2 cm^3), and the solution stirred for a further 30 min. The reaction was then quenched by the addition of 0.5 M TEAB buffer (10 cm^3), and stored at 4 °C overnight. The solution was evaporated to dryness and redissolved in water (20 cm^3) and applied to a Sephadex A25 column in 0.05 M TEAB buffer. The column was eluted with a linear gradient of 0.05−1.0 M TEAB. Appropriate fractions were pooled and evaporated to dryness to give a yellow solid of methyl-1-(2-deoxy-β-D-ribofuranosyl)-5-nitroindole-3-carboxylate 5′-triphosphate. Yield 110 mg. HPLC (Phenomenex Luna 10 μ C-18 reverse phase column, buffer A, 0.1 M TEAB; buffer B, 0.1 M TEAB, 25% MeCN. Twenty-five to 100% buffer B over 45 min at 8 mL/min) showed the product to be pure. δ_P (D_2O) −9.35 (d, γ-P), −10.15 (d, α-P), −22.05 (t, β-P). A solution of methyl-1-(2-deoxy-β-D-ribofuranosyl)-5-nitroindole-3-carboxylate-5′-triphosphate (70 mg) in 0.880 ammonia (10 cm^3) was stirred at room temperature overnight. HPLC showed complete conversion. The solution was evaporated to a yellow solid, and the product purified by HPLC. δ_P (D_2O) −5.15 (d, γ-P), −10.10 (d, α-P), −21.25 (t, β-P). The title compound was converted into its sodium salt by passage through a Dowex 50WX4−200 resin (Na^+ form). Yield 418.8 OD. δ_P (D_2O) −6.92 (d, γ-P), −9. (d, α-P), −21.05 (t, β-P).

Selection, Screening, Protein Expression and Characterization. For selection we used the previously described library 3T (1×10^9 cfu, 70% active clones).[40] Emulsification and CSR selection were performed as described[31,70] using primers 1, 2 for rounds 1, 2, primers 1−4 for round 3 and 3−7 for rounds 4−5, cycled 20× (94 °C 30 s, 50 °C 30 s, 72 °C 5 min), reamplified with gene specific primers 8−13 for rounds 1, 2, with primers 8−13 and out-nested primers 14, 15 or combinations thereof for rounds 3−5 and recloned *Xba* I/*Sal*I into pASK75 as described.[31] After selection rounds one and two, clones were screened by d5NIC PCR with primers 1,2 or 3,4 and by polymerase ELISA as described[63] using hairpins 16−19 or 20−23. Promising clones from rounds 3 and 4 were StEP shuffled[71] and backcrossed with parent polymerase genes. Clones analyzed in more detail in this report derive from selection rounds 4 and 5. Selected mutations were reverted back to wild-type Taq sequence using Quickchange Mutagenesis using Pfu Turbo (Stratagene) and primers 31−40. Expression of polymerases for characterization was as described[40] using a 16/10 Hi-Prep Heparin FF Column (Amersham Pharmacia Biotech). Polymerase fractions eluted around 0.3 M NaCl and were concentrated and dia-filtered into 50 mM Tris pH 7.4, 1 mM DTT, 50% glycerol and stored at −20 °C. Mutation rates were determined using a well-established *in vivo* gap filling assay[69] (M. Arana, PH, T. Kunkel, unpublished results). 5D4 PCR products with primers 1−7 and pASK75 as template, were reamplified using 5D4 with primers 14, 15, TOPO cloned (Invitrogen) and sequenced.

Primer Extension and PCR Assays. Extension reactions with purified polymerases were carried out by addition of 4 μL of 2.5 mM dNTP mix (final concentration 50 μM each dATP, dTTP, dCTP, dGTP) to 46 μL of a reaction mixture of final concentration containing 1× Taq buffer; 50 pmol ^{32}P-labeled primers; 20, 21 or 28, 100 pmol templates; 22−26, 29, 30, and 1 μL of polymerase (wtTaq (1.5 μg) or 5D4 (16 μg), activity normalized) at 60 °C. Eight microliter aliquots were removed and added to 8 μL of stop solution (8 M urea, 50 mM EDTA, ∼0.1% xylene cyanol F) at 0, 10, 20, 30, and 40 min, and the products were electrophoretically separated on 20% polyacrylamide gels. Kinetic primer extension reactions were carried out by mixing equal volumes of primers 20 or 21/templates 22−25 (100 μL stock solution containing 1× Taq buffer, 80 pmol ^{32}P-labeled primer and 200 pmol template) and polymerase (wtTaq or 5D4)/dXTP mix (100 μL stock solution containing 1× Taq buffer, dXTP to final concentration between

[65] Lu, H.; He, K.; Kool, E. T. *Angew. Chem., Int. Ed.* **2004**, *43*, 5834–5836.

[66] Bebenek, K.; Joyce, C. M.; Fitzgerald, M. P.; Kunkel, T. A. *J. Biol. Chem.* **1990**, *265*, 13878–13887.

[67] Loakes, D.; Guo, M. J.; Brown, D. M.; Salisbury, S. A.; Smith, C. L.; Felix, I. R.; Kumar, S.; Nampalli, S. *Nucleosides, Nucleotides Nucleic Acids* **2000**, *19*, 1599–1614.

[68] Loakes, D.; Hill, F.; Brown, D. M.; Ball, S.; Reeve, M. A.; Robinson, P. S. *Nucleosides, Nucleotides Nucleic Acids* **1999**, *18*, 2685–2695.

[69] Pon, R. T. *Protocols for oligonucleotides and analogs*; Humana Press: Totowa, NJ, 1993; Vol. 20, p 469.

[70] Ghadessy, F. J.; Holliger, P. *Methods Mol. Biol.* **2007**, *352*, 237–248.

[71] Zhao, H.; Giver, L.; Shao, Z.; Affholter, J. A.; Arnold, F. H. *Nat. Biotechnol.* **1998**, *16*, 258–261.

1−160 μM and polymerase, X = dA, dT, dC, dA, d5NI, d5NIC). Reaction mixtures were mixed at 60 °C and quenched after various time intervals by the addition of an equal volume of stop solution (8 M urea, 50 mM EDTA, ~0.1% xylene cyanol F) before electrophoretic separation on 20% polyacylamide gel. Kinetic reactions were all performed in triplicate. kcat/Km values are in %μM^{-1} min^{-1}. Polyacrylamide gels were dried and exposed to a phosphorimager screen (Amersham Biosciences or Molecular Dynamics) and scanned on a Typhoon 8610 (Molecular Dynamics). Data was initially analyzed using Geltrak[72,73] and processed using Kaleidagraph (Synergy Software) or Excel (Microsoft). PCR assays were performed using primers 1−7, pASK75 template and PCR conditions 20× (94 °C 30 s, 50 °C 30 s, 72 °C 30 s) for d5NIC primers and 50× (94 °C 30 s, 50 °C 30 s, 72 °C 5 min) for d5NI primers on a MJ TETRAD thermocycler.

NMR Spectroscopy and NMR Model Calculation. NMR spectra of *tni* and *tnic* (Figure 11) were acquired on Bruker DRX-500 and DMX-600 spectrometers, processed using NMRPIPE,[74] and analyzed using Sparky 3.106.[75] Two-dimensional NMR spectra recorded in D_2O included $^1H-^{31}P$ HetCOR and uninterrupted series of dqf-COSY, TOCSY, ROESY and NOESY (with 60, 120 and 250 ms mixing times) experiments. NOESY spectra were also acquired in H_2O at 9 °C with a mixing time of 120 ms. For structure determination, distance restraints were estimated from NOESY build-ups using NOE interactions corresponding to covalently constrained interproton distances as a reference; base-pair hydrogen-bonding restraints were introduced based on chemical shifts and interactions observed in H_2O-NOESY experiments; and sugar−phosphate backbone dihedral restraints were deduced from the cross-peak patterns observed in dqf-COSY, $^1H-^{31}P$ HetCOR and NOESY spectra.[50] Based on strong NMR evidence (Figure 4, Supporting Information), the C6-G13 double-helical stems (Figure 10A) of *tni* and *tnic* were constrained to a B-DNA conformation, and NMR models of *tni* and *tnic* were calculated by restrained molecular dynamics, using extensive distance and dihedral NMR restraint sets for the T1-C6 and G13-d5NI(C)15 nucleotides (Supporting Table NMR), a distance-dependent dielectric constant, and the MMFF94 force field[76] of SYBYL 6.9 (Tripos Inc.).

Statistical Coupling Analysis. Starting from the Conserved Domain Database[77] entry on the polymerase A family (cd06444), 3987 protein sequences were identified in the NCBI database as putative PolA by the position-specific score matrix of the family. Of those, 3484 were of bacterial origin and were selected for further processing. Clustal W[78] and manual curation were used to identify duplicate entries and to reduce sampling bias due to single species overrepresentation. The resulting data set contained 1057 sequences, from which sequence alignment, using MUSCLE,[79] was carried out iteratively to remove sequences of dubious quality and to trim the alignment to the polymerase domain. The resulting sequence set, used in the statistical coupling analysis, contained 994 polymerase domain sequences aligned to the *Thermus aquaticus* DNA polymerase A domain (E432−E832 in PDB strucutre 3KTQ[80]) (available on request).

SCA[55] was performed using the SCA toolbox v3.0, kindly provided by R. Ranganathan (Dallas), in MATLAB (The Mathworks, Inc.) using the 994-polymerase alignment. SCA was carried out using *Thermus aquaticus*, *Geobacillus stearothermophilus* and *Escherichia coli* as query sequences and structures. The SCA output correlation matrix (available on request) was further analyzed using Excel. An arbitrary identity cutoff of 97% was used for analysis. Aligned residues conserved above the cutoff were not considered in the analysis as subalignments used for SCA calculations for those residues would contain fewer than 30 sequences. The log-normal fit of the SCA correlations suggested 0.85 kT^* (mean plus 3 SDs) as the significance threshold. We initially set an arbitrary 1.80 kT^* cutoff to identify the most relevant couplings. A more detailed analysis of the 5D4 mutations was also carried out using the 0.85 kT^* cutoff to identify couplings from 5D4 mutations with greater sensitivity.

Protein Stability Analysis Using FoldX. FoldX[61] (version 3.0beta3, http://foldx.crg.es/foldx.jsp) was used to predict the effect of mutations found in 5D4 on protein stability by comparing the free energy of folding between mutants and wild-type Taq. Comparisons were carried out using 1CMW[81] (apo structure of full-length Taq), 1KTQ[82] (apo structure of Klentaq fragment), 2KTQ[80] (Klentaq fragment in an open ternary complex) and 3KTQ[80] (Klentaq fragment in a closed ternary complex), and the effect of insertion of individual 5D4 mutations into the Taq framework was computed.

Acknowledgment. J.G. is grateful for the support of an EU Marie Curie Fellowship (QLK2-CT-1999-51436). We thank Prof. R. Ranganathan (Green Centre for Systems Biology, Dallas, TX) for providing SCA toolbox v3.0. E.T.K. acknowledges grants from the National Institutes of Health (GM072705, GM067201 and GM63587).

Supporting Information Available: Protein sequences of selected polymerases (including 5D4), additional data on kinetic constants, substrate specificity, primer extension, NMR spectroscopy, reversion mutant activity, SCA and FoldX analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

JA9039696

(72) Smith, J. M.; Thomas, D. J. *Comput. Appl. Biosci.* **1990**, *6*, 93–99.
(73) Smith, J.; Singh, M. *BioTechniques* **1996**, *20*, 1082–1087.
(74) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277–293.
(75) Goddard, T. D.; Kneller, D. G., ; University of California: San Francisco, CA, 2001. http://www.cgl.ucsf.edu/home/sparky/.
(76) Halgren, T. A. *J. Comp. Chem.* **1996**, *17*, 490–519.
(77) Marchler-Bauer, A.; Anderson, J. B.; Chitsaz, F.; Derbyshire, M. K.; DeWeese-Scott, C.; Fong, J. H.; Geer, L. Y.; Geer, R. C.; Gonzales, N. R.; Gwadz, M.; He, S.; Hurwitz, D. I.; Jackson, J. D.; Ke, Z.; Lanczycki, C. J.; Liebert, C. A.; Liu, C.; Lu, F.; Lu, S.; Marchler, G. H.; Mullokandov, M.; Song, J. S.; Tasneem, A.; Thanki, N.; Yamashita, R. A.; Zhang, D.; Zhang, N.; Bryant, S. H. *Nucleic Acids Res.* **2009**, *37*, D205–210.
(78) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.
(79) Edgar, R. C. *BMC Bioinformatics* **2004**, *5*, 113.
(80) Li, Y.; Korolev, S.; Waksman, G. *EMBO J.* **1998**, *17*, 7514–7525.
(81) Urs, U. K.; Murali, R.; Krishna Murthy, H. M. *Acta Crystallogr. D: Biol. Crystallogr.* **1999**, *55*, 1971–1977.
(82) Korolev, S.; Nayal, M.; Barnes, W. M.; Di Cera, E.; Waksman, G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9264–9268.