ACCOUNTS
of chemical research

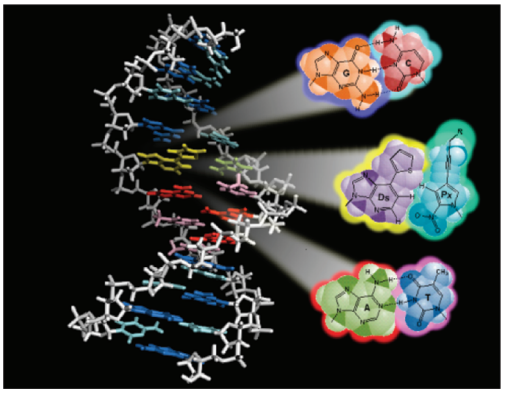
Natural versus Artificial Creation of Base Pairs in DNA: Origin of Nucleobases from the Perspectives of Unnatural Base Pair Studies

ICHIRO HIRAO,*,†,‡ MICHIKO KIMOTO,†,‡ AND RIE YAMASHIGE†

†*RIKEN Systems and Structural Biology Center (SSBC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, and* ‡*TagCyx Biotechnologies, 1-6-126 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan*

RECEIVED ON OCTOBER 6, 2011

CONSPECTUS



Since life began on Earth, the four types of bases (A, G, C, and T(U)) that form two sets of base pairs have remained unchanged as the components of nucleic acids that replicate and transfer genetic information. Throughout evolution, except for the U to T modification, the four base structures have not changed. This constancy within the genetic code raises the question of how these complicated nucleotides were generated from the molecules in a primordial soup on the early Earth. At some prebiotic stage, the complementarity of base pairs might have accelerated the generation and accumulation of nucleotides or oligonucleotides. We have no clues whether one pair of nucleobases initially appeared on the early Earth during this process or a set of two base pairs appeared simultaneously.

Recently, researchers have developed new artificial pairs of nucleobases (unnatural base pairs) that function alongside the natural base pairs. Some unnatural base pairs in duplex DNA can be efficiently and faithfully amplified in a polymerase chain reaction (PCR) using thermostable DNA polymerases. The addition of unnatural base pair systems could expand the genetic alphabet of DNA, thus providing a new mechanism for the generation novel biopolymers by the site-specific incorporation of functional components into nucleic acids and proteins. Furthermore, the process of unnatural base pair development might provide clues to the origin of the natural base pairs in a primordial soup on the early Earth. In this Account, we describe the development of three representative types of unnatural base pairs that function as a third pair of nucleobases in PCR and reconsider the origin of the natural nucleic acids.

As researchers developing unnatural base pairs, they use repeated "proof of concept" experiments. As researchers design new base pairs, they improve the structures that function in PCR and eliminate those that do not. We expect that this process is similar to the one functioning in the chemical evolution and selection of the natural nucleobases. Interestingly, the initial structures designed by each research group were quite similar to those of the latest successful unnatural base pairs. In this regard, it is tempting to form a hypothesis that the base pairs on the primordial Earth, in which the natural purine bases, A and G, and pyrimidine bases, C and T(U), originated from structurally similar compounds, such as hypoxanthine for a purine base predecessor. Subsequently, the initial base pair evolved to the present two sets of base pairs via a keto-enol tautomerization of the initial compounds.

Introduction

The complementary A–T(U) and G–C base pairs in nucleic acids are central to life on Earth. These bases function as the alphabet of genetic information in DNA and RNA, and their structures are considered to have remained unchanged during the evolution process after a certain stage of the RNA world. Theoretical approaches have speculated that the optimized number of base types is four for replicative genetic information storage,[1,2] although a computational analysis also suggested the possibilities of six and eight.[3]

Published on the Web 01/20/2012 www.pubs.acs.org/accounts
10.1021/ar200257x

If that is the case, then when and how did the four different bases and their complementarity appear in a prebiotic stage on the early Earth? In other words, is it even possible to simultaneously generate the very complicated ribonucleotides of the four base types? Some bases and nucleotides were obtained from quite primitive molecules under conditions simulating the primordial Earth,[4,5] and pyrimidine ribonucleotides were also generated under prebiotically plausible conditions.[6] An evolution experiment indicated that two bases, which pair with each other, are sufficient for generating a ligase ribozyme, although the activity was low.[7] Yet, we still have no clue as to how the complementarity originally appeared in the forms of the A–T and G–C pairs.

For the last two decades, the development of artificial third base pairs (unnatural base pairs) has been pursued for exploring genetic alphabet expansion with replicable DNA molecules (Figure 1).[8–12] Researchers have thus created several types of unnatural base pairs that function in replication with high fidelity and efficiency, in combination with the natural A–T and G–C pairs. These unnatural base pairs were created by repeating design processes on the basis of certain ideas – chemical synthesis of their nucleotides and oligonucleotides, and performing physical and biological assays to assess the selectivity and efficiency of the base pair formation. The promising base pairs were improved by modifying their constituents, and the others were eliminated. This developmental process resembles the process observed in natural generation (chemical evolution) and selection. Interestingly, the structures of the successful unnatural base pairs with replication fidelity are quite similar to those of their initial candidates.

When comparing the process of unnatural base pair development with that of natural base pair creation on Earth, we cannot help but imagine the origin of the A–T(U) and G–C pairs, which might be similar to the original ones that appeared as the first complementary molecule set on the primordial Earth. Here, we will describe the development process of three representative types of unnatural base pairs that function as a third base pair in PCR, and reconsider the origin of the natural base pairs. Although many valuable reports about modified bases and base analogues exist, these are not the subject of this Account.

Design of Unnatural Base Pairs for Replication

Alexander Rich imagined a new base pair system including a third base pair in 1962,[1] and pioneering studies of unnatural

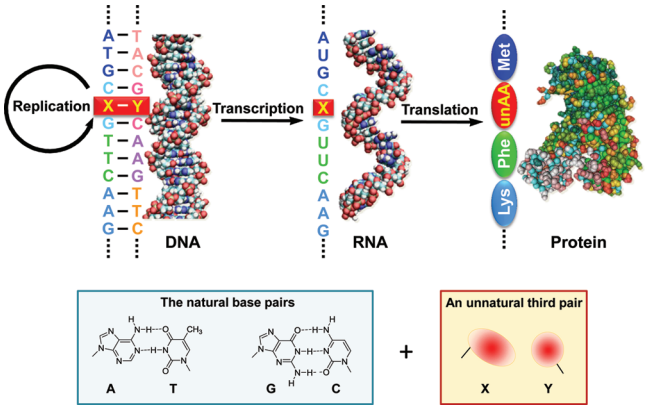


FIGURE 1. Expansion of the genetic alphabet by an unnatural base pair (**X**–**Y**) system.

base pairs were begun in the late 1980s.[13,14] Since then, many candidates were designed, chemically synthesized, and tested for their complementarity in a biology system. In order to expand the genetic alphabet, the unnatural base pairs must have highly exclusive selectivity, orthogonality, in which the fifth base always pairs with the sixth one, as in the A–T and G–C pairings, in the DNA duplex and the template-directed nucleic acid synthesis by polymerases (Figure 1). The latest unnatural base pairs can be practically used in PCR amplification as a third base pair, and their selectivity is higher than 99.8% per replication or PCR cycle.

In 1990, Benner's group designed four types of unnatural base pairs with nonstandard hydrogen-bonding patterns,[15] which differed from those of the natural base pairs. Although the ability of their early unnatural base pairs was not high, due to tautomerization and chemical instability,[16] they overcame these problems to develop a new base pair with the same concept.[17,18] Meanwhile, Kool's group synthesized nonhydrogen-bonded base pairs between shape-analogues of the natural bases,[19] and suggested the importance of shape complementarity, for base pairing. Although Kool's base pairs are isosteres of the natural base pairs and cannot be used as a third base pair, their studies inspired researchers further. Subsequently, nonhydrogen-bonded hydrophobic base analogues also became candidates for unnatural base pairs. Romesberg's group synthesized many hydrophobic base analogues,[20] and finally found optimized unnatural base pairs for PCR.[21,22] Our group has also developed highly specific hydrophobic base pairs,[23] through consecutive improvements by combining the concepts of nonstandard hydrogen bonding patterns, shape-complementarity, hydrophobicity, and electrostatic repulsion, on an ongoing basis.[24–28] These developmental processes are described in more depth below.

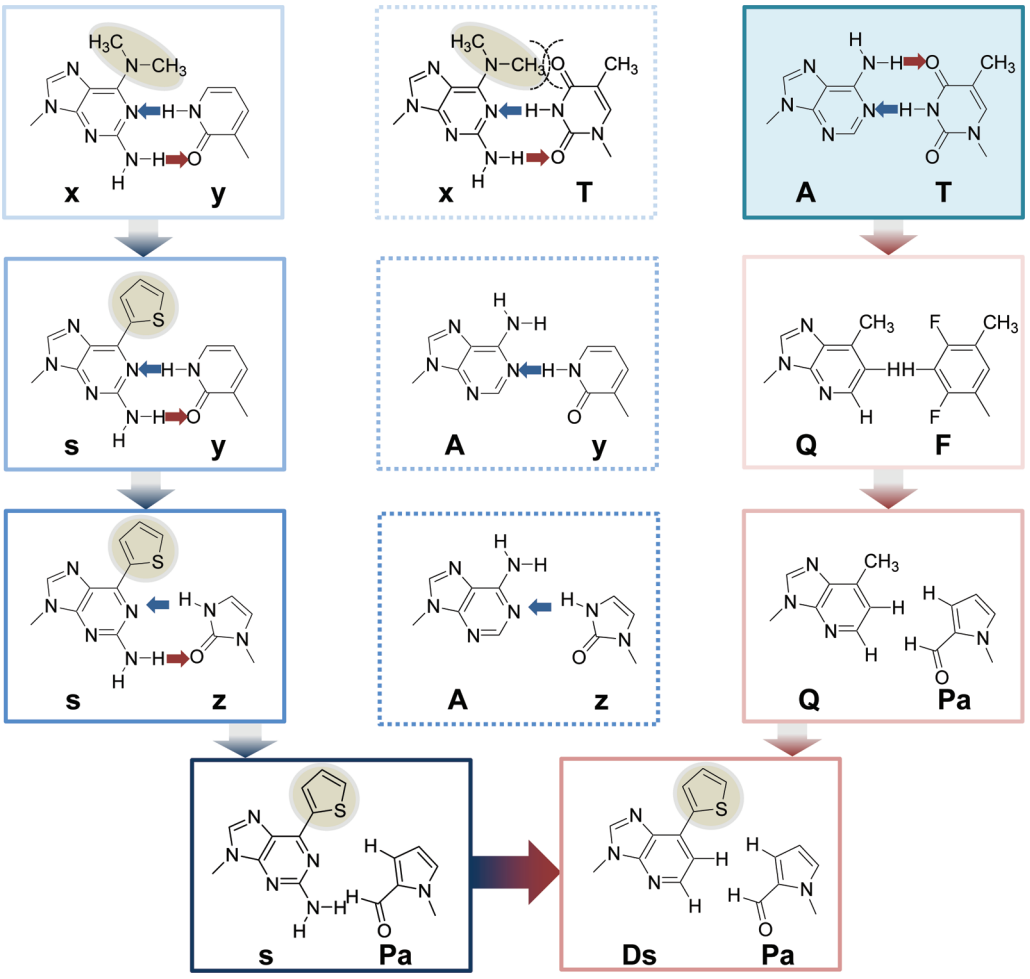


FIGURE 2. Hirao's unnatural base pair development process: from the **x**–**y** and **Q**–**Pa** pairs to the **Ds**–**Pa** pair. The **Q**–**F** pair was synthesized by Kool's group as a hydrophobic A–T pair analogue.

Hirao's Hydrophobic Ds–Px Pair

In 2000, we developed our first unnatural base pair, between 2-amino-6-dimethylaminopurine (**x**) and pyridine-2-one (**y**) (Figure 2),[24] by combining two concepts, a nonstandard hydrogen bonding pattern and shape-complementarity. The 2-aminopurine part of **x** may pair with T with two hydrogen bonds, and to avoid this **x**–T mispairing we introduced a bulky dimethylamino group to **x**, which clashes with the 4-keto group of T (Figure 2). In addition, we included a hydrogen atom in **y**, to accommodate the dimethylamino group in the **x**–**y** pair, which thus exhibited specificity in transcription. T7 RNA polymerase incorporates the ribonucleoside triphosphate of **y** (**y**TP) into RNA opposite **x** in DNA templates with more than 90% selectivity.[24] However, the **x**–**y** pair in replication, using the 3′–5′ exonuclease-deficient Klenow fragment of *Escherichia coli* DNA polymerase I (KF exo$^-$), was not selective (see Figure 4).

To increase the steric hindrance of the bulky group of **x**, we replaced the flexible dimethylamino group with a highly planar thienyl group, and synthesized 2-amino-6-(2-thienyl)-purine (**s**) as a pairing partner of **y** (Figure 2).[25] In addition to the steric hindrance of **s**, the sulfur atom in the thienyl group electrostatically clashes with the 4-keto group of T. The **s**–**y** pair showed higher selectivity than the **x**–**y** pair in both replication (Figure 4) and transcription. In transcription, T7 RNA polymerase incorporates **y**TP into RNA opposite **s**, with more than 96% selectivity. This **s**–**y** pair transcription was combined with an in vitro translation system, allowing the site-specific incorporation of a nonstandard amino acid into a protein.[25]

However, the selectivity of the **s**–**y** pair is still insufficient for PCR. In the **s**–**y** pair, **s** has a bulky thienyl group to exclude mispairing with any natural bases, in contrast to **y**. Thus, besides d**s**TP, the natural purine substrates, especially dATP, are also incorporated opposite **y** in replication (Figure 2).

Actually, after 10 cycles of PCR, most of the **s**–**y** pairs were replaced with the A–T pairs in the amplified DNA.

To address this problem, we employed a five-membered-ring structure, imidazolin-2-one (**z**), instead of the six-membered-ring of **y** (Figure 2).[29] The shape complementarity between the larger **s** and smaller **z** bases was improved, relative to that of the **s**–**y** pair. In addition, the fitting between A and **z** seemed to be loose (Figure 2), making it difficult to eliminate the solvating waters on these bases for the A–**z** mispair formation. In transcription, the **s**–**z** selectivity was greatly improved, as compared to that of the **s**–**y** pair. However, the low hydrophobicity and poor stacking ability of **z** weakened the interaction with polymerases, causing reduced incorporation efficiency of d**z**TP into DNA. Thus, we shifted our attention to considering hydrophobicity and stacking interactions, besides shape complementarity, for base pair design.

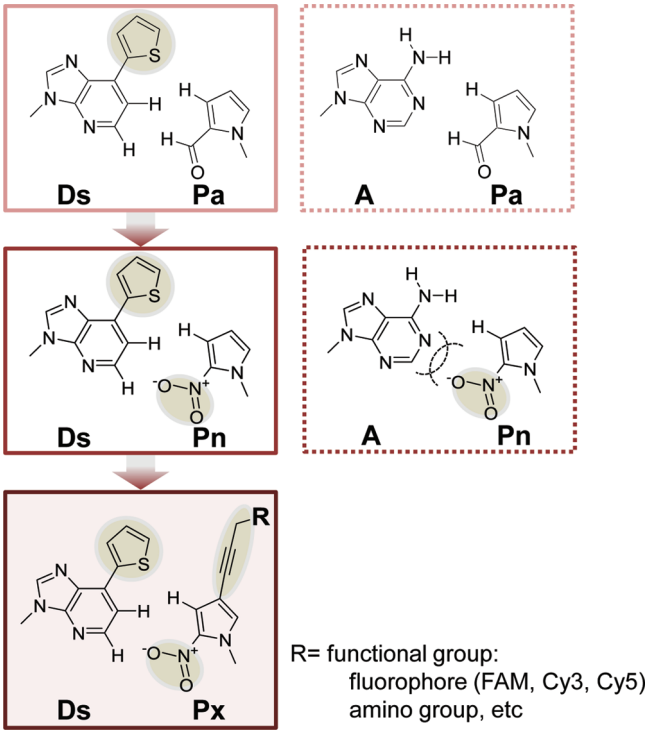


FIGURE 3. Hirao's unnatural **Ds**–**Pa**, **Ds**–**Pn**, and **Ds**–**Px** base pairs, which function in PCR.

In the mid-1990s, Kool's group reported hydrophobic A–T pair analogues, such as the **Z**–**F**[19] and **Q**–**F** pairs[30] (Figure 2). These unnatural base pairs lack hydrogen-bonding interactions between the pairing bases, but function in replication. The **Q** base has a nitrogen atom corresponding to position 3 in the natural purines, for interactions with the side chains of amino acids in polymerases. While considering the **Q**–**F** pair, we noticed that it could be improved by employing a five-membered-ring base analogue, pyrrole-2-carboaldehyde (**Pa**),[26] as a pairing partner of **Q**, instead of the six-membered-ring of the **F** base (Figure 2). The shape of **Pa** fits better with **Q**, in comparison to **F**, and the oxygen atom of the aldehyde group increases the interaction with polymerases. The **Q**–**Pa** pair exhibited higher selectivity than that of the **Q**–**F** pair in replication (Figure 4). The tertiary structures of each DNA duplex containing the **Q**–**Pa**[26] or **Z**–**F**[31] pair were determined by NMR (Figure 5). The shape complementarity of **Pa** with **Q** is better than that of **F** with **Z**, and the base pair structure of the **Q**–**Pa** pair in the duplex is more geometrically similar to those of the natural base pairs. Furthermore, we accidentally found that **Pa** also pairs with **s**, and the **s**–**Pa** pair functions in transcription with higher efficiency than the precedent **s**–**z** pair (Figure 2).[32,33]

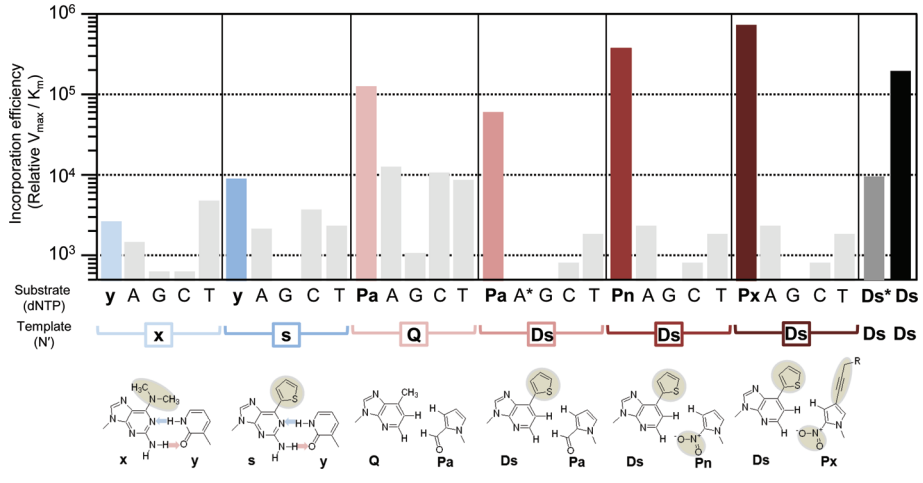


FIGURE 4. Single-nucleotide incorporation efficiency and selectivity of a series of Hirao's unnatural base pairs by the 3′–5′ exonuclease-deficient Klenow fragment of *E. coli* DNA polymerase I. **Ds*** and A* are their γ-amidotriphosphates.

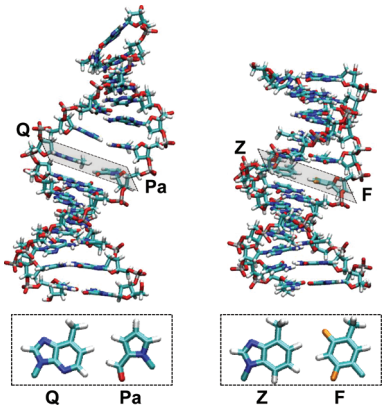


FIGURE 5. Structures of DNA duplexes containing the **Q**–**Pa** or **Z**–**F** pair, determined by NMR. The geometry of each unnatural base pair moiety in the duplex is shown at the bottom.

Based on the **s**–**Pa** and **Q**–**Pa** pairs, we subsequently removed the hydrogen-bonding residues from the **s** base or replaced the methyl group of **Q** with a bulky thienyl moiety, and created hydrophobic 7-(2-thienyl)imidazo[4,5-*b*]-pyridine (**Ds**),[27] as a new pairing partner of **Pa** (Figure 2). The hydrogen-bonding residues of **s** still facilitate mispairing with the natural bases, and **Q** pairs with T, as well as with **Pa**. Thus, we employed the thienyl group, which efficiently excludes the mispairing with T, like that in **s**. The **Ds**–**Pa** pair functions complementarily in transcription and replication. However, there is another problem with **Ds**–**Ds** mispairing in replication: the incorporation efficiency of d**Ds**TP opposite **Ds** is higher than that of d**Pa**TP opposite **Ds** (Figure 4). Large hydrophobic bases, such as **Ds**, easily stack on each other in the DNA duplex, and replication at the site pauses due to the structural distortion.

We overcame this problem by using modified triphosphate substrates, γ-amidotriphosphates, which can reduce the mispairing between bases with less shape-complementarity in replication.[27] We thus developed a highly specific replication system involving the **Ds**–**Pa** pair, using the combination of the γ-amidotriphosphates of **Ds** and A, and the usual triphosphates of **Pa**, G, C, and T. The γ-amidotriphosphates of A efficiently exclude the mispairing between A and **Pa**. DNA fragments containing the **Ds**–**Pa** pair can be amplified by PCR using Vent DNA polymerase (exo$^+$) with an unnatural base pair selectivity of 99.0% per replication, and approximately 96–97% of the **Ds**–**Pa** pair remained in the amplified DNA fragments after 20 cycles of PCR.

Our next task was to develop unnatural base pairs without the help of the γ-amidotriphosphates, which reduce PCR amplification efficiency and limit further applications. Thus, we focused on the improvement of the **Pa** base. Instead of **Pa**, we searched for a new base that more efficiently pairs with **Ds** and avoids the **Ds**–**Ds** and A–**Pa** mispairings. To exclude the A–**Pa** pair, we replaced the aldehyde group of **Pa** with a nitro group, and designed 2-nitropyrrole (**Pn**).[28] The oxygen atom of the nitro group was expected to electrostatically repel the 1-nitrogen of **A** (Figure 3). Indeed, **Pn** was efficiently and selectively incorporated opposite **Ds** and greatly reduced the occurrence of mispairing with A. DNA fragments containing the **Ds**–**Pn** pair can be amplified ~500-fold with >99.0% selectivity by 20 cycles of PCR, with only the γ-amidotriphosphate of **Ds**, not that of A.

Furthermore, to exclude the **Ds**–**Ds** mispairing, we increased the hydrophobicity of the **Pn** base and thus designed 2-nitro-4-propynylpyrrole(**Px**), by attaching a propynyl group to **Pn** (Figure 3).[23] The increased hydrophobicity strengthened the interaction with polymerases and the stacking with neighboring bases, resulting in the superior incorporation efficiency of d**Px**TP opposite **Ds**, as compared to that of d**Ds**TP opposite **Ds** (Figure 4). Without any modified triphosphates, the **Ds**–**Px** pair exhibited high efficiency and fidelity in PCR. DNA fragments containing the **Ds**–**Px** pair can be amplified to 10^7-fold by 30 cycles of PCR using Deep Vent DNA polymerase (exo$^+$), and the **Ds**–**Px** pairing selectivity was over 99.9%. Furthermore, a variety of functional groups, such as a fluorophore and biotin, can be attached to the amino group of **Px**, and these functional d**Px**TPs can also be site-specifically incorporated into DNA by replication. Recently, we confirmed that the **Ds**–**Px** pair can survive through 100 cycles of PCR, and more than 97% of the **Ds**–**Px** pair in DNA was retained in the 10^{28}-fold amplified products after 100-cycle PCR (10-cycle PCR repeated 10 times).[34]

Romesberg's Hydrophobic 5SICS–MMO2 Pair

In 1999, Romesberg's group reported the predominantly hydrophobic self-pair of propynyl isocarbostyril (**PICS**), as their first successful unnatural base pair (Figure 6).[20] They showed the high thermal stability of the **PICS**–**PICS** pair in DNA duplexes and the site-specific incorporation of d**PICS**TP opposite **PICS** in replication by KF exo$^-$. These results indicated that hydrophobic packing increases the duplex stability and the incorporation efficiency. Subsequently, they chemically synthesized a wide variety of hydrophobic unnatural base analogues and evaluated their base pairing ability in replication with several DNA polymerases.[35–37]

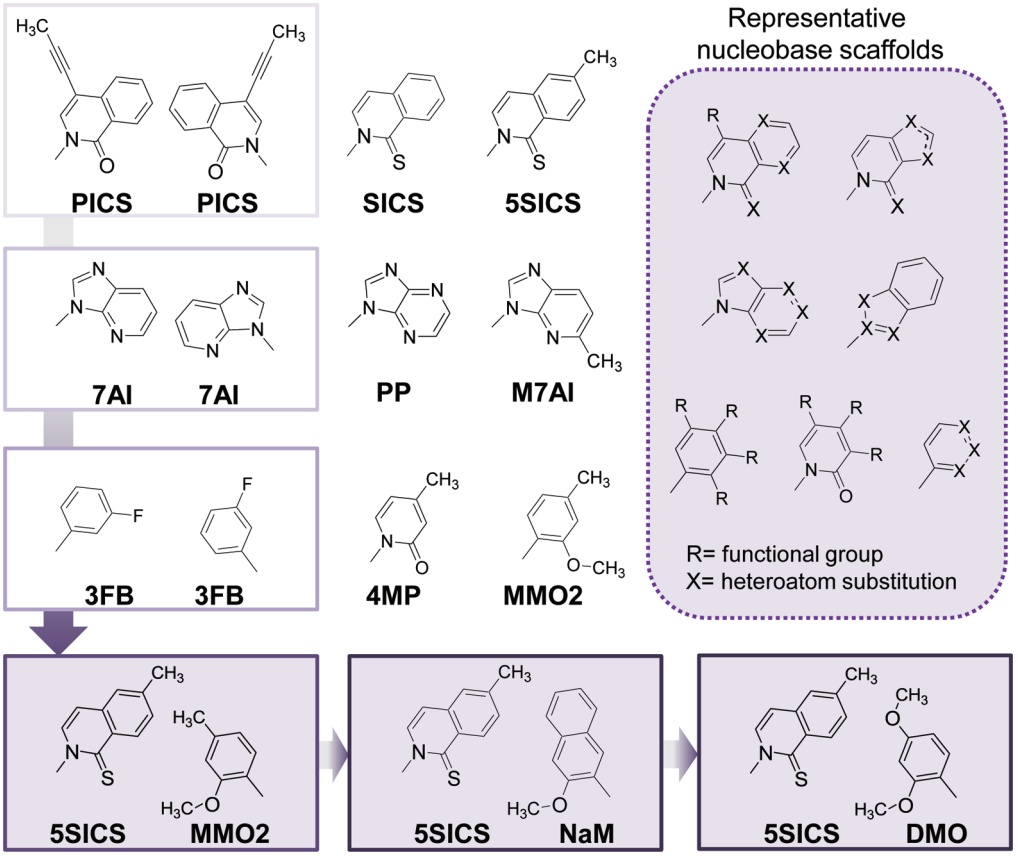


FIGURE 6. Romesberg's hydrophobic unnatural bases and base pairs. The **5SICS**–**MMO2**, **5SICS**–**NaM**, and **5SICS-DMO** pairs function in PCR.

They found that the unnatural base pairs between relatively large hydrophobic base analogs, such as the **PICS**–**PICS** pair and the **7AI**–**7AI** pair[36] (Figure 6), exhibit high incorporation efficiency.

However, a problem with these unnatural base pairs was the poor extension after their incorporation. They analyzed the tertiary structure of the **PICS**–**PICS** pair in a DNA duplex by NMR, and found that the **PICS** bases partially stack on each other. The stacked structure facilitates the unnatural base incorporation, but terminates further extension because of the distorted stacking structure, which was similar to the **Ds**–**Ds** mispairing, as mentioned above. They overcame this issue by using two polymerases, the Klenow fragment for the single-nucleotide insertion and rat pol β for further extension,[38] as well as by mutating the polymerase by an in vivo evolution method, using the proteolytic fragment of *Taq* (Stoffel fragment).[39]

Simultaneously, they extensively investigated a variety of base pair combinations comprising relatively small hydrophobic base analogs for replication extension.[40−43] A phenyl ring analog with a single fluorine substituent, 3-fluorobenzene (**3FB,** Figure 6),[43] was efficiently incorporated self-complementarily, and further primer extension occurred after the **3FB**–**3FB** pairing by KF exo$^-$. Their structural and biochemical studies revealed that the **3FB**–**3FB** pair in the DNA duplex forms a naturally planar pair structure, suitable for primer extension.[44]

Romesberg's group synthesized more than 60 base analogues (representatives are shown in Figure 6). Thus, they applied a combinatorial approach by two independent screening methods, using 3600 (60 $\times$ 60) possible hydrophobic base pairs, to identify candidates for efficient replication.[45] All possible base pair combinations were evaluated by single-nucleotide insertion experiments and further extension by KF exo$^-$. Among the combinations, the **SICS**–**MMO2** pair exhibited the best efficiency and selectivity. Furthermore, they also investigated the structure–activity relationships of the designed base pairs, and optimized the structures by small modifications.[45,46]

They finally developed the **5SICS**–**MMO2**, **5SICS**–**NaM**, and **5SICS**–**DMO** pairs (Figure 6),[21,22] which all function in PCR. The methyl group of **5SICS** reduces the **5SICS**–**5SICS** self-mispairing.[45] **NaM** was generated by introducing a second aromatic ring to **MMO2**, fused at the *meta* and

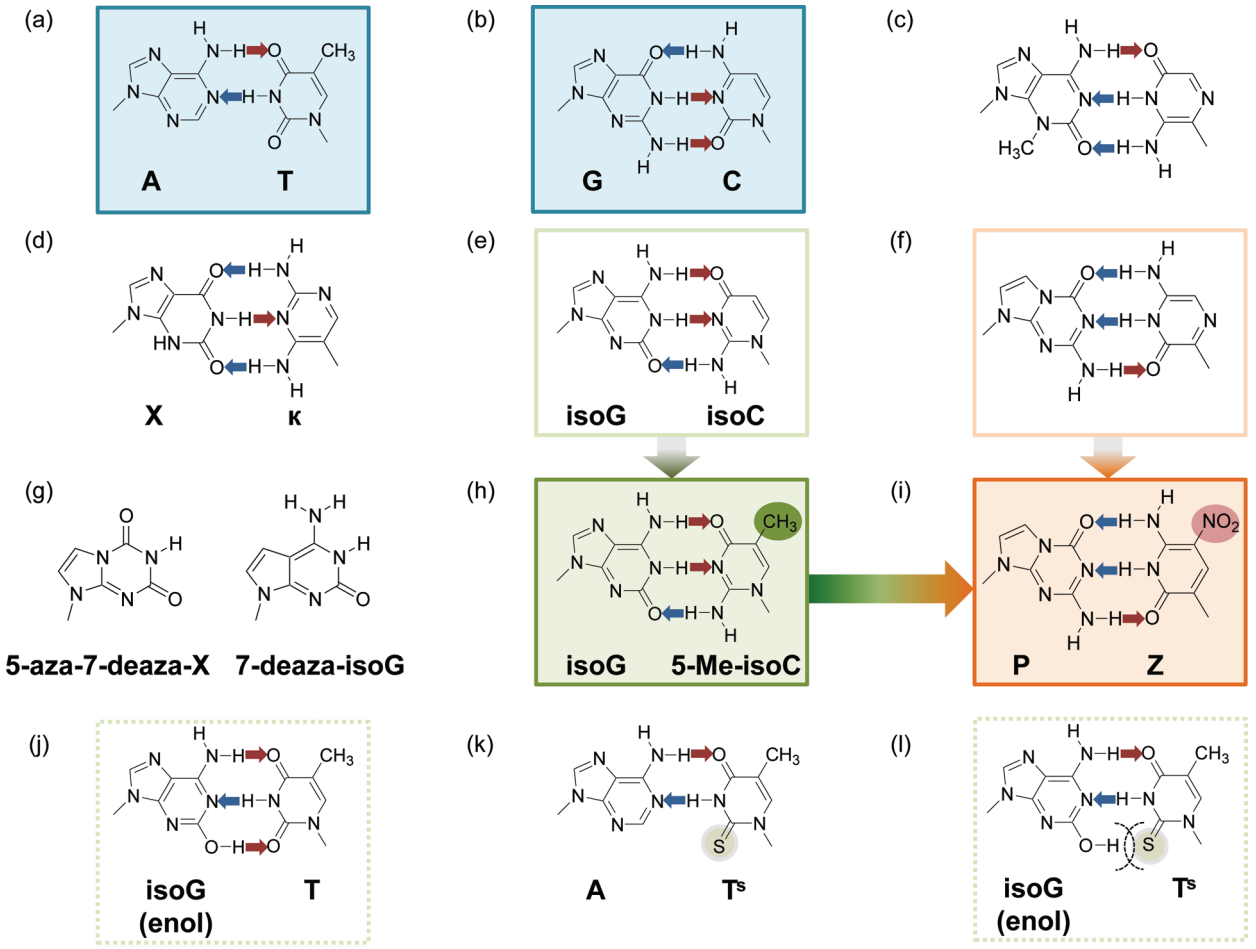


FIGURE 7. Natural A–T and G–C pairs and four additional unnatural base pairs with different hydrogen-bonding patterns proposed by Benner's group (a–f). Nucleobase analogues for xanthine and isoG (g). The improved **isoG**–5-methyl-**isoC** and **P**–**Z** pairs function in PCR (h,i). The **isoG** enol tautomer pairs with T (j). The A–T^s pair replacing the A–T pair, reduces **isoG** (enol) –T^s mispairing (k,l).

para positions.[46] In addition, in **DMO**, the *para*-methyl substituent of **MMO2** was replaced with a methoxy group by fine-tuning the shape-complementarity,[22] and the d**DMO**TP incorporation efficiency opposite **5SCIS**, rather than **MMO2**, was increased. They also determined the structure of the **5SICS**–**MMO2** pair in a DNA duplex, which revealed that the unnatural pairing bases are slightly stacked on each other.[22] They proposed that this unique, slight stacking causes the high specificity of the **5SICS**–**MMO2** pairing, in which the subtle balance between the stacked and unstacked structures is the key for efficient replication. The **5SICS**–**MMO2**, **5SICS**–**NaM**, and **5SICS**–**DMO** pairs function well in PCR using Deep Vent DNA polymerase. The fidelities per replication cycle ranged from 92.9–99.4% for **5SICS**–**MMO2**, 98.0–99.8% for **5SICS**–**NaM**, and 90.7–99.8% for **5SICS**–**DMO** in PCR amplification. The **5SICS**–**MMO2** and **5SICS**–**NaM** pairs can also function in transcription using T7 RNA polymerase.[47]

Benner's Hydrogen-Bonded Z–P Pair

The initial development of unnatural base pairs by Benner's group was based on the rational design of different hydrogen bonding patterns from those of the natural Watson–Crick base pairs (Figure 7). In 1989, they experimentally demonstrated that the isoguanine (**isoG**) and isocytosine (**isoC**) pair (Figure 7e)[1] can be incorporated into DNA and RNA in replication and transcription.[13] Benner's group also synthesized another unnatural base pair, xanthine(**X**) and 2,6-diaminopyrimidine (*κ*), to further extend the concept (Figure 7a–f).[5] In 1992, they reported a new codon–anticodon combination including the **isoG**–**isoC** pair, which functions in in vitro translation to synthesize a peptide containing a nonstandard amino acid, using chemically synthesized mRNA and tRNA.[48]

However, these pioneering studies were hampered by various shortcomings of the base pairs. First, the **isoC**, *κ*, and **X** bases have proton-donor amino or imino groups, instead

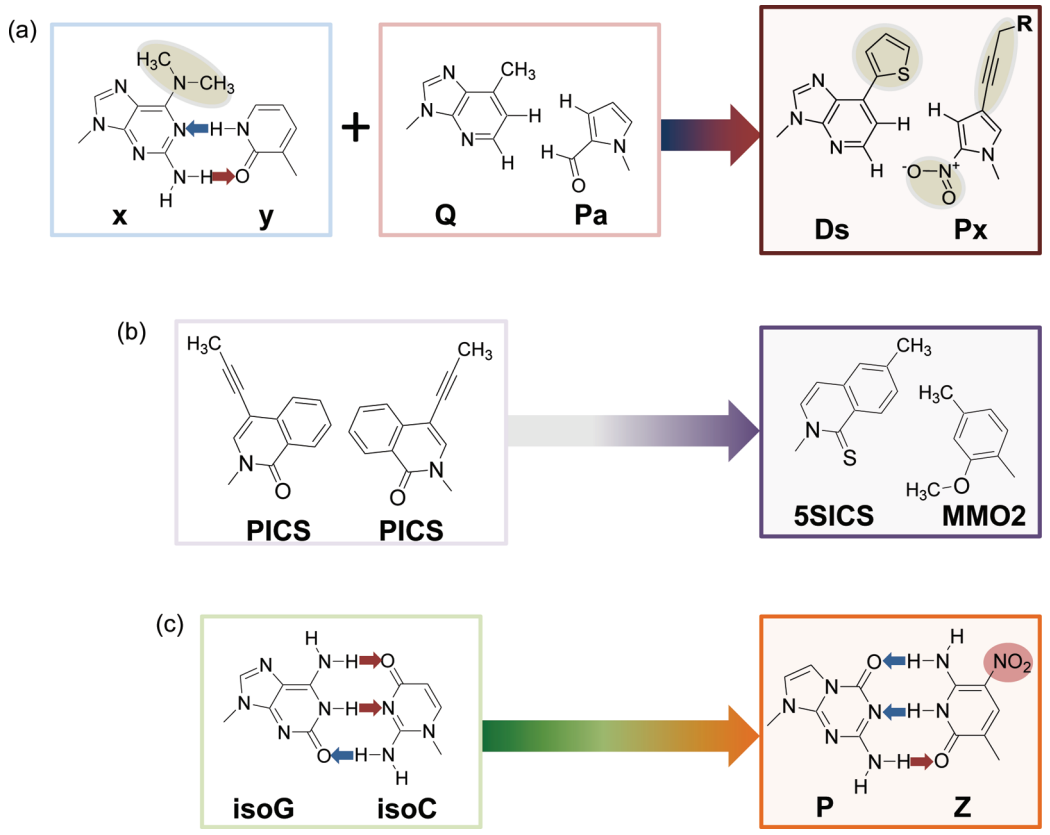


FIGURE 8. Successful unnatural base pairs and their origins.

of proton-acceptor residues for interaction with polymerases, decreasing their incorporation efficiencies.[16,49] Second, **X** is anionic under physiological conditions ($pK_a \sim 5.6$),[50] destabilizing DNA duplexes.[8] This problem was managed by employing 5-aza-7-deazaxanthine (Figure 7g).[51] Third, the **isoC** (Figure 7e) and pyrazine-scaffold-base (Figure 7f) nucleosides are chemically unstable, and easily decompose under mild alkaline conditions.[8,13] Subsequent analyses revealed that 5-position substituents, such as 5-methyl-**isoC** (Figure 7h), increased the stability.[52,53] Lastly, the keto–enol tautomerization of **isoG** under physiological pH conditions is a serious problem.[16,54] About 10% of the enol tautomer of **isoG** is present in an aqueous solution, and it pairs with T(U) (Figure 7j),[55] causing the low fidelity of the **isoG**–**isoC** pair, which is mutated to the A–T pair during PCR amplification.

To solve the tautomerism problem, Benner's group employed two strategies. One is the use of 7-deaza-**isoG** (Figure 7g),[55] which favors the keto form over the enol form (ca. 1000:1). Another solution is the use of a thymine analogue, 2-thio-T(T^s), instead of T (Figure 7k).[56] The 2-thione group of T^s reduces its hydrogen bonding ability, decreasing the pairing with the enol form of **isoG** (Figure 7l). Using the three types of base pairs, **isoG**–5-methyl-**isoC**, A–T^s, and G–C, the unnatural base pair selectivity reached ~98% in PCR by Titanium*Taq* DNA polymerase.

Their latest unnatural base pair is that between 2-aminoimidazo[1,2-*a*]-1,3,5-triazin-4(8*H*)-one (**P**) and 6-amino-5-nitro-2(1*H*)-pyridone (**Z**) (Figure 7i).[17,18,57] Both bases have proton-acceptor residues in the minor groove side, to enhance their interactions with the polymerase. The **P** and **Z** bases are now free from tautomerism and chemical instability. The nitropyridine scaffold of **Z** resists oxidization and epimerization. They recently reported the high fidelity (99.8% per theoretical PCR cycle) of the **P**–**Z** pairing in PCR using *Taq* DNA polymerase.[18]

Natural versus Artificial Chemical Evolution of Base Pairs

Here, we have introduced the developmental process for three types of unnatural base pairs that function in PCR with high selectivity. The hydrophobic **Ds**–**Px** pair was generated by a consecutive improvement process from two initial key base pairs, hydrogen-bonded **x**–**y** and hydrophobic **Q**–**Pa** pairs (Figure 8a). The hydrophobic **5SICS**–**MMO2** pair was derived from the initial hydrophobic **PICS**–**PICS**

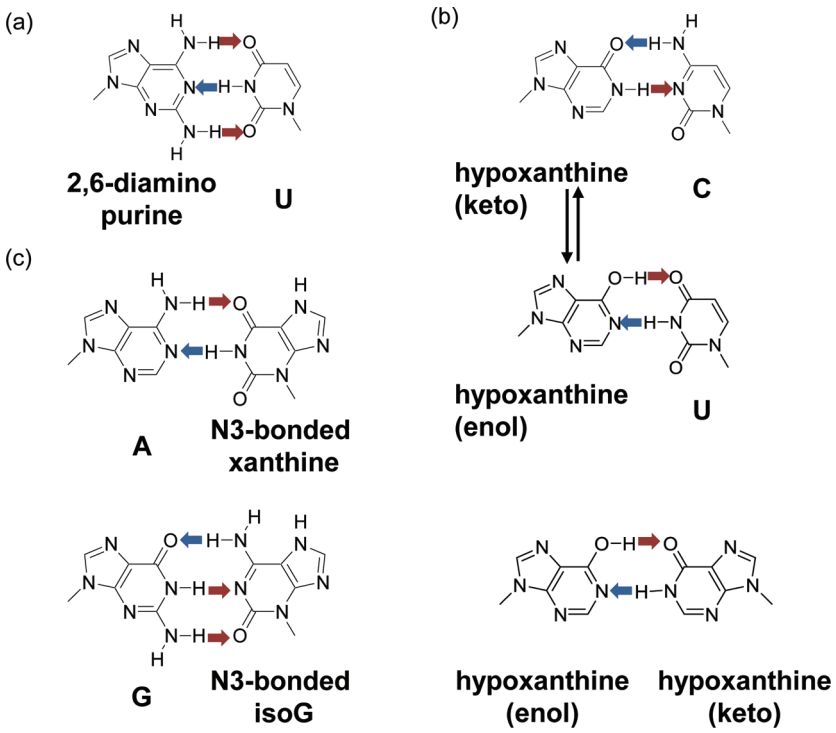


FIGURE 9. Candidate bases for the origin of the natural base pairs. The 2,6-diaminopurine–uracil base pair (a). Hypoxanthine tautomers can pair with U, C, and hypoxanthine (b). Base pairs proposed by Wächtershäuser (c).

pair via rational selection from combinatorial libraries (Figure 8b). The hydrogen-bonded **P**–**Z** pair was created by solving several problems with the initial **isoG**–**isoC** pair (Figure 8c). Although each of these successful base pairs was designed using different ideas and concepts, the chemical structures of these initial base pairs are quite similar to those of the latest ones (Figure 8). From this perspective of artificial chemical evolution, we consider the origin of the A–T(U) and G–C pairs on the primordial Earth.

When visualizing the structures of the A–T(U) and G–C pairs, besides the complementarity of the pairing bases in each pair, we notice the symmetry of the hydrogen-bonding patterns between A–T(U) and G–C. This symmetric property suggests that the natural base pairs may have been generated from one type of initial base or base pair, with a similar structure to the present one. In a chemical reaction, the structural symmetry of two molecules is generated from one molecule by a reaction, such as racemization and optical inactivation. In addition, the idea that fewer than four base types appeared at the beginning easy explains the prebiotic synthesis of the quite complicated components, ribonucleotides. Several researchers hypothesized that two bases would be sufficient,[1,58,59] and Joyce's group experimentally demonstrated that oligoribonucleotides consisting of only one type of base pair, between 2,6-diaminopurine and uracil (Figure 9a), function as a ligase ribozyme.[7]

An alternative initial candidate for the predecessor of the A–T(U) and G–C pairs might be hypoxanthine, capable of pairing with both T(U) and C by tautomerism (Figure 9b).[60] Thus, the intrinsic symmetric property of hydrogen-bond patterns appears in hypoxanthine between the keto and enol forms, as shown in the **isoG** base tautomerization. Although the keto form of hypoxanthine predominates over the enol form in a neutral solution and a wobbling pair between hypoxanthine and U is also plausible, the higher acidity ($pK_a = 8.8$) of the proton at position N3 of hypoxanthine relative to that ($pK_a = 9.2$) of guanine indicates its tautomerism.[61,62] As possible pairing partners of hypoxanthine, the nucleotide derivatives of C and U were generated under primordial-like conditions.[6] Hypoxanthine was also experimentally obtained from very primitive molecules under conditions simulating a primordial earth,[5] and A and G were derived from hypoxanthine by amination reactions. Another precursor of the base pairs is a hypoxanthine self-pair between its keto and enol forms (Figure 9b). A similar hypothesis was proposed by Wächtershäuser for adenine–N3-bonded xanthine and guanine–N3-bonded isoguanine pairs (Figure 9c).[63] Either way, the symmetric

hydrogen bonding patterns of A and G strongly resemble those of the keto and enol forms of hypoxanthine.

Once the A−T(U) and G−C pairs appeared on the primordial Earth, they remained invariant until now as nucleic acid components. In contrast to this immortality of the intrinsic natural base pair structures, researchers have been developing further unique unnatural base pairs. Kool's group studies new genetic systems using unnatural bases with expanded sizes.[64] Matsuda and Minakawa's group developed hydrophilic unnatural base pairs with four hydrogen bonds.[65] Recently, Carell's group created a metal-salen base pair capable of reversible bond formation, enabling PCR amplification.[66] Thus, unnatural base pairs are still in midstream development in artificial chemical evolution.

We thank Bor Hodošček and our collaborators for contributions and stimulating discussions. This work was supported by Grants-in-Aid for Scientific Research (KAKENHI 19201046 to I.H., 20710176 to M.K.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Targeted Proteins Research Program and the RIKEN Structural Genomics/Proteomics Initiative, the National Project on Protein Structural and Functional Analyses, Ministry of Education, Culture, Sports, Science and Technology of Japan.

BIOGRAPHICAL INFORMATION

Ichiro Hirao is the team leader of the nucleic acid synthetic biology research team in RIKEN, Systems and Structural Biology Center, and the CEO of TagCyx Biotechnologies. He earned his Ph.D. in Chemistry from Tokyo Institute of Technology. His interests include a wide range of research areas of biopolymers, focusing on nucleic acids from chemistry to biology.

Michiko Kimoto earned her Ph.D. at the University of Tokyo in 2002, and subsequently joined Hirao's group. Since 2006, she has been a research scientist in RIKEN and TagCyx Biotechnologies.

Rie Yamashige earned her master's degree at the University of Miyazaki in 2008, and joined Hirao's group in 2009. Since 2011, she has been a research associate in RIKEN.

FOOTNOTES

*To whom correspondence should be addressed. Fax +81-45-503-9645. E-mail ihirao@riken.jp.

REFERENCES

1 Rich, A. Problems of evolution and biochemical information transfer. In *Horizons in Biochemistry*; Kasha, M. P. B., Ed.; Academic Press: 1962; pp 103−126.

2 Szathmary, E. Why are there four letters in the genetic alphabet? *Nat. Rev. Genet.* **2003**, *4*, 995–1001.

3 Gardner, P. P.; Holland, B. R.; Moulton, V.; Hendy, M.; Penny, D. Optimal alphabets for an RNA world. *Proc. Biol. Sci.* **2003**, *270*, 1177–1182.

4 Oro, J. Mechanism of synthesis of adenine from hydrogen cyanide under possible primitive earth conditions. *Nature* **1961**, *191*, 1193–1194.

5 Sanchez, R. A.; Ferris, J. P.; Orgel, L. E. Studies in prebiotic synthesis. IV. Conversion of 4-aminoimidazole-5-carbonitrile derivatives to purines. *J. Mol. Biol.* **1968**, *38*, 121–128.

6 Powner, M. W.; Gerland, B.; Sutherland, J. D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **2009**, *459*, 239–242.

7 Reader, J. S.; Joyce, G. F. A ribozyme composed of only two different nucleotides. *Nature* **2002**, *420*, 841–844.

8 Benner, S. A. Understanding nucleic acids using synthetic chemistry. *Acc. Chem. Res.* **2004**, *37*, 784–797.

9 Henry, A. A.; Romesberg, F. E. Beyond A, C, G and T: augmenting nature's alphabet. *Curr. Opin. Chem. Biol.* **2003**, *7*, 727–733.

10 Krueger, A. T.; Kool, E. T. Redesigning the architecture of the base pair: toward biochemical and biological function of new genetic sets. *Chem. Biol.* **2009**, *16*, 242–248.

11 Hirao, I. Unnatural base pair systems for DNA/RNA-based biotechnology. *Curr. Opin. Chem. Biol.* **2006**, *10*, 622–627.

12 Kimoto, M.; Cox, R. S., 3rd; Hirao, I. Unnatural base pair systems for sensing and diagnostic applications. *Expert Rev. Mol. Diagn.* **2011**, *11*, 321–331.

13 Switzer, C.; Moroney, S. E.; Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.* **1989**, *111*, 8322–8323.

14 Rappaport, H. P. The 6-thioguanine/5-methyl-2-pyrimidinone base pair. *Nucleic Acids Res.* **1988**, *16*, 7253–7267.

15 Piccirilli, J. A.; Krauch, T.; Moroney, S. E.; Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **1990**, *343*, 33–37.

16 Switzer, C. Y.; Moroney, S. E.; Benner, S. A. Enzymatic recognition of the base pair between isocytidine and isoguanosine. *Biochemistry* **1993**, *32*, 10489–10496.

17 Yang, Z.; Sismour, A. M.; Sheng, P.; Puskar, N. L.; Benner, S. A. Enzymatic incorporation of a third nucleobase pair. *Nucleic Acids Res.* **2007**, *35*, 4238–4249.

18 Yang, Z.; Chen, F.; Alvarado, J. B.; Benner, S. A. Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *J. Am. Chem. Soc.* **2011**, *133*, 15105–15112.

19 Morales, J. C.; Kool, E. T. Efficient replication between non-hydrogen-bonded nucleoside shape analogs. *Nat. Struct. Biol.* **1998**, *5*, 950–954.

20 McMinn, D. L.; Ogawa, A. K.; Wu, Y.; Liu, J.; Schultz, P. G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: DNA polymerase recognition of a highly stable, self-pairing hydrophobicbase. *J. Am. Chem. Soc.* **1999**, *121*, 11585–11586.

21 Malyshev, D. A.; Seo, Y. J.; Ordoukhanian, P.; Romesberg, F. E. PCR with an expanded genetic alphabet. *J. Am. Chem. Soc.* **2009**, *131*, 14620–14621.

22 Malyshev, D. A.; Pfaff, D. A.; Ippoliti, S. I.; Hwang, G. T.; Dwyer, T. J.; Romesberg, F. E. Solution structure, mechanism of replication, and optimization of an unnatural base pair. *Chemistry* **2010**, *16*, 12650–12659.

23 Kimoto, M.; Kawai, R.; Mitsui, T.; Yokoyama, S.; Hirao, I. An unnatural base pair system for efficient PCR amplification and functionalization of DNA molecules. *Nucleic Acids Res.* **2009**, *37*, e14.

24 Ohtsuki, T.; Kimoto, M.; Ishikawa, M.; Mitsui, T.; Hirao, I.; Yokoyama, S. Unnatural base pairs for specific transcription. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4922–4925.

25 Hirao, I.; Ohtsuki, T.; Fujiwara, T.; Mitsui, T.; Yokogawa, T.; Okuni, T.; Nakayama, H.; Takio, K.; Yabuki, T.; Kigawa, T.; Kodama, K.; Nishikawa, K.; Yokoyama, S. An unnatural base pair for incorporating amino acid analogs into proteins. *Nat. Biotechnol.* **2002**, *20*, 177–182.

26 Mitsui, T.; Kitamura, A.; Kimoto, M.; To, T.; Sato, A.; Hirao, I.; Yokoyama, S. An unnatural hydrophobic base pair with shape complementarity between pyrrole-2-carbaldehyde and 9-methylimidazo[(4,5)-b]pyridine. *J. Am. Chem. Soc.* **2003**, *125*, 5298–5307.

27 Hirao, I.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Kawai, R.; Sato, A.; Harada, Y.; Yokoyama, S. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. *Nat. Methods* **2006**, *3*, 729–735.

28 Hirao, I.; Mitsui, T.; Kimoto, M.; Yokoyama, S. An efficient unnatural base pair for PCR amplification. *J. Am. Chem. Soc.* **2007**, *129*, 15549–15555.

29 Hirao, I.; Harada, Y.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Yokoyama, S. A two-unnatural-base-pair system toward the expansion of the genetic code. *J. Am. Chem. Soc.* **2004**, *126*, 13298–13305.

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

31 Guckian, K. M.; Krugh, T. R.; Kool, E. T. Solution structure of a DNA duplex containing a replicable difluorotoluene-adenine pair. *Nat. Struct. Biol.* **1998**, *5*, 954–959.

32 Kimoto, M.; Mitsui, T.; Harada, Y.; Sato, A.; Yokoyama, S.; Hirao, I. Fluorescent probing for RNA molecules by an unnatural base-pair system. *Nucleic Acids Res.* **2007**, *35*, 5360–5369.

33 Hikida, Y.; Kimoto, M.; Yokoyama, S.; Hirao, I. Site-specific fluorescent probing of RNA molecules by unnatural base-pair transcription for local structural conformation analysis. *Nat. Protoc.* **2010**, *5*, 1312–1323.

34 Yamashige, R.; Kimoto, M.; Takezawa, Y.; Sato, A.; Mitsui, T.; Yokoyama, S.; Hirao, I. Highly specific unnatural base pair systems as a third base pair for PCR amplification. *Nucleic Acids Res.* DOI:10.1093/nar/GKR1068.

35 Ogawa, A. K.; Wu, Y.; McMinn, D. L.; Liu, J.; Schultz, P. G.; Romesberg, F. E. Efforts toward the expansion of the genetic alphabet: Information storage and replication with unnatural hydrophobic base pairs. *J. Am. Chem. Soc.* **2000**, *122*, 3274–3287.

36 Wu, Y.; Ogawa, A. K.; Berger, M.; McMinn, D. L.; Schultz, P. G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: Optimization of interbase hydrophobic interactions. *J. Am. Chem. Soc.* **2000**, *122*, 7621–7632.

37 Ogawa, A. K.; Wu, Y.; Berger, M.; Schultz, P. G.; Romesberg, F. E. Rational design of an unnatural base pair with increased kinetic selectivity. *J. Am. Chem. Soc.* **2000**, *122*, 8803–8804.

38 Tae, E. L.; Wu, Y.; Xia, G.; Schultz, P. G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: replication of DNA with three base pairs. *J. Am. Chem. Soc.* **2001**, *123*, 7439–7440.

39 Leconte, A. M.; Chen, L.; Romesberg, F. E. Polymerase evolution: efforts toward expansion of the genetic code. *J. Am. Chem. Soc.* **2005**, *127*, 12470–12471.

40 Berger, M.; Luzzi, S. D.; Henry, A. A.; Romesberg, F. E. Stability and selectivity of unnatural DNA with five-membered-ring nucleobase analogues. *J. Am. Chem. Soc.* **2002**, *124*, 1222–1226.

41 Henry, A. A.; Olsen, A. G.; Matsuda, S.; Yu, C.; Geierstanger, B. H.; Romesberg, F. E. Efforts to expand the genetic alphabet: identification of a replicable unnatural DNA self-pair. *J. Am. Chem. Soc.* **2004**, *126*, 6923–6931.

42 Leconte, A. M.; Matsuda, S.; Romesberg, F. E. An efficiently extended class of unnatural base pairs. *J. Am. Chem. Soc.* **2006**, *128*, 6780–6781.

43 Matsuda, S.; Leconte, A. M.; Romesberg, F. E. Minor groove hydrogen bonds and the replication of unnatural base pairs. *J. Am. Chem. Soc.* **2007**, *129*, 5551–5557.

44 Matsuda, S.; Fillo, J. D.; Henry, A. A.; Rai, P.; Wilkens, S. J.; Dwyer, T. J.; Geierstanger, B. H.; Wemmer, D. E.; Schultz, P. G.; Spraggon, G.; Romesberg, F. E. Efforts toward expansion of the genetic alphabet: structure and replication of unnatural base pairs. *J. Am. Chem. Soc.* **2007**, *129*, 10466–10473.

45 Leconte, A. M.; Hwang, G. T.; Matsuda, S.; Capek, P.; Hari, Y.; Romesberg, F. E. Discovery, characterization, and optimization of an unnatural base pair for expansion of the genetic alphabet. *J. Am. Chem. Soc.* **2008**, *130*, 2336–2343.

46 Seo, Y. J.; Hwang, G. T.; Ordoukhanian, P.; Romesberg, F. E. Optimization of an unnatural base pair toward natural-like replication. *J. Am. Chem. Soc.* **2009**, *131*, 3246–3252.

47 Seo, Y. J.; Matsuda, S.; Romesberg, F. E. Transcription of an expanded genetic alphabet. *J. Am. Chem. Soc.* **2009**, *131*, 5046–5047.

48 Bain, J. D.; Switzer, C.; Chamberlin, A. R.; Benner, S. A. Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature* **1992**, *356*, 537–539.

49 Horlacher, J.; Hottiger, M.; Podust, V. N.; Hubscher, U.; Benner, S. A. Recognition by viral and cellular DNA polymerases of nucleosides bearing bases with nonstandard hydrogen bonding patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6329–6333.

50 Kulikowska, E.; Kierdaszuk, B.; Shugar, D. Xanthine, xanthosine and its nucleotides: solution structures of neutral and ionic forms, and relevance to substrate properties in various enzyme systems and metabolic pathways. *Acta. Biochim. Pol.* **2004**, *51*, 493–531.

51 Rao, P.; Benner, S. A. Fluorescent charge-neutral analogue of xanthosine: synthesis of a 2′-deoxyribonucleoside bearing a 5-aza-7-deazaxanthine base. *J. Org. Chem.* **2001**, *66*, 5012–5015.

52 Roberts, C.; Bandaru, R.; Switzer, C. Theoretical and experimental study of isoguanine and isocytosine: Base pairing in an expanded genetic system. *J. Am. Chem. Soc.* **1997**, *119*, 4640–4649.

53 Tor, Y.; Dervan, P. B. Site-specific enzymatic incorporation of an unnatural base, N6-(6-aminohexyl)isoguanosine, into RNA. *J. Am. Chem. Soc.* **1993**, *115*, 4461–4467.

54 Sepiol, J.; Kazimierczuk, Z.; Shugar, D. Tautomerism of isoguanosine and solvent-induced keto-enol equilibrium. *Z. Naturforsch.* **1976**, *31*, 361–370.

55 Martinot, T. A.; Benner, S. A. Artificial genetic systems: exploiting the ″aromaticity″ formalism to improve the tautomeric ratio for isoguanosine derivatives. *J. Org. Chem.* **2004**, *69*, 3972–3975.

56 Sismour, A. M.; Benner, S. A. The use of thymidine analogs to improve the replication of an extra DNA base pair: a synthetic biological system. *Nucleic Acids Res.* **2005**, *33*, 5640–5646.

57 Yang, Z.; Chen, F.; Chamberlin, S. G.; Benner, S. A. Expanded genetic alphabets in the polymerase chain reaction. *Angew. Chem., Int. Ed.* **2010**, *49*, 177–180.

58 Crick, F. H. C. Origin of Genetic Code. *J. Mol. Biol.* **1968**, *38*, 367–379.

59 Orgel, L. E. Evolution of the genetic apparatus. *J. Mol. Biol.* **1968**, *38*, 381–393.

60 Shugar, D.; Kierdaszuk, B. New Light on Tautomerism of Purines and Pyrimidines and Its Biological and Genetic-Implications. *J. Biosci.* **1985**, *8*, 657–668.

61 Chenon, M. T.; Pugmire, R. J.; Grant, D. M.; Panzica, R. P.; Townsend, L. B. Carbon-13 magnetic resonance. XXVI. A quantitative determination of the tautomeric populations of certain purines. *J. Am. Chem. Soc.* **1975**, *97*, 4636–4642.

62 Burkard, M. E.; Turner, D. H. NMR structures of r(GCAGGCGUGC)2 and determinants of stability for single guanosine-guanosine base pairs. *Biochemistry* **2000**, *39*, 11748–11762.

63 Wachtershauser, G. An all-purine precursor of nucleic acids. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 1134–1135.

64 Liu, H.; Gao, J.; Lynch, S. R.; Saito, Y. D.; Maynard, L.; Kool, E. T. A four-base paired genetic helix with expanded size. *Science* **2003**, *302*, 868–871.

65 Ogata, S.; Takahashi, M.; Minakawa, N.; Matsuda, A. Unnatural imidazopyridopyrimidine: naphthyridine base pairs: selective incorporation and extension reaction by Deep Vent(exo$^-$) DNA polymerase. *Nucleic Acids Res.* **2009**, *37*, 5602–5609.

66 Kaul, C.; Muller, M.; Wagner, M.; Schneider, S.; Carell, T. Reversible bond formation enables the replication and amplification of a crosslinking salen complex as an orthogonal base pair. *Nat. Chem.* **2011**, *3*, 794–800.