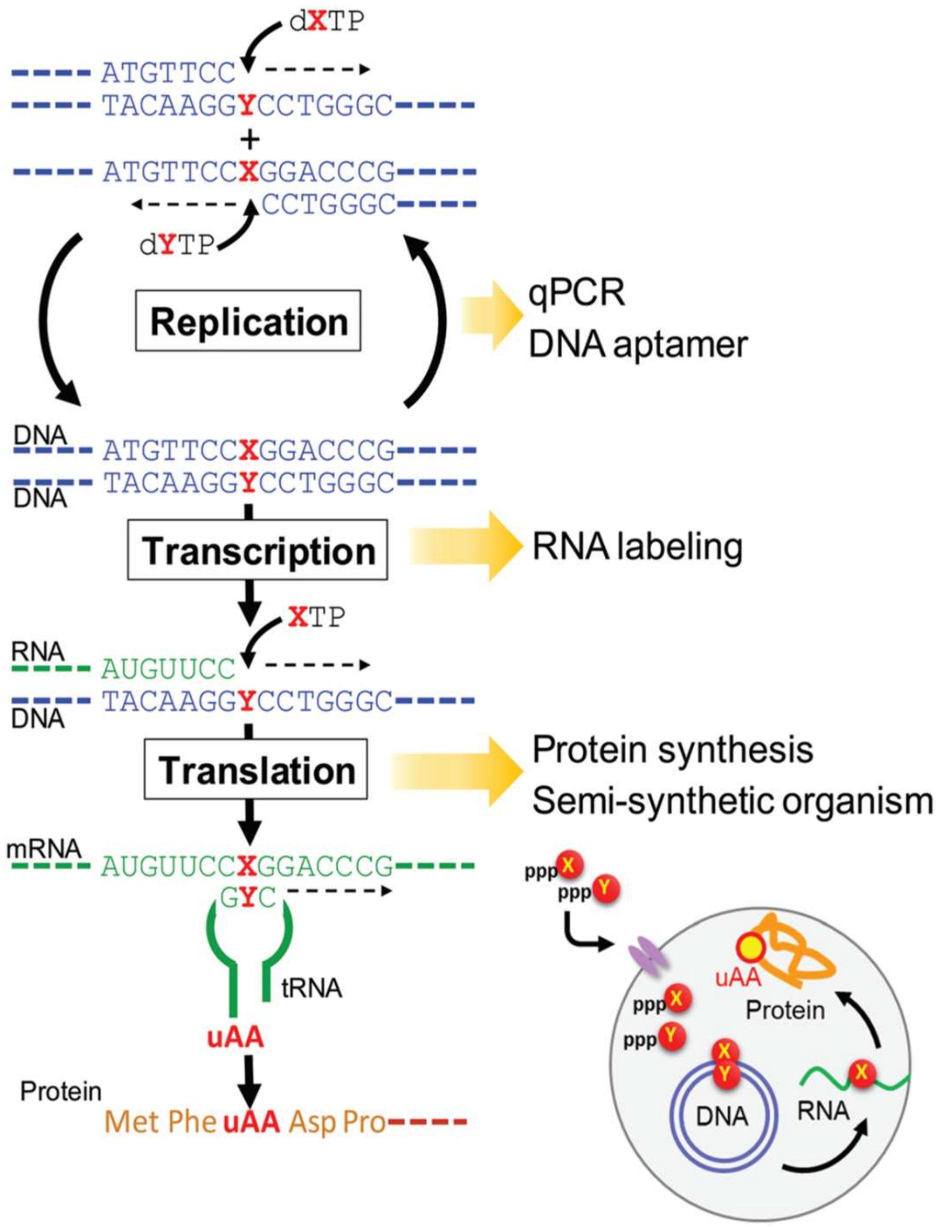
Expansion of the genetic alphabet

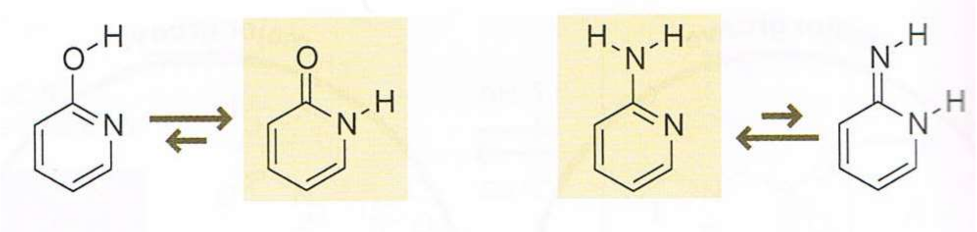
Expansion of the genetic alphabet and code by creating an unnatural base pair (UBP) as a third pair.

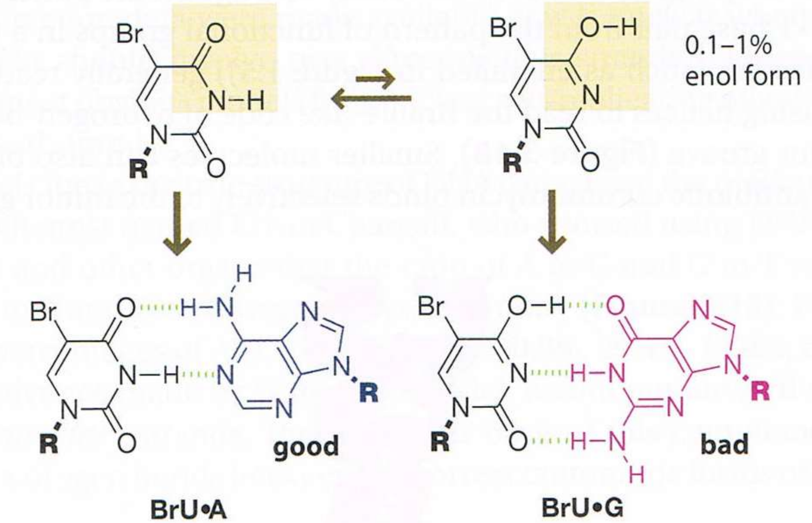
The creation of a UBP (i.e., X–Y) that functions in replication, transcription, and translation as a third base pair with the natural A–T(U) and G–C pairs allows the storage and retrieval of the expanded genetic information in vitro and in vivo, enabling a variety of applications using biopolymers with increased functionalities

M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626



Tautomeric instability in non-canonical nucleobases





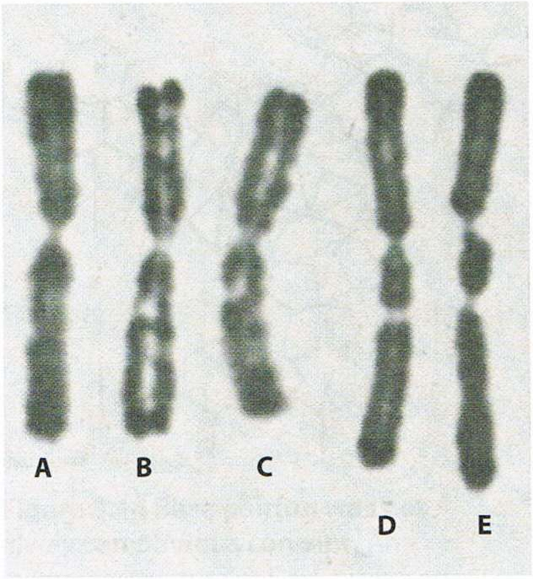
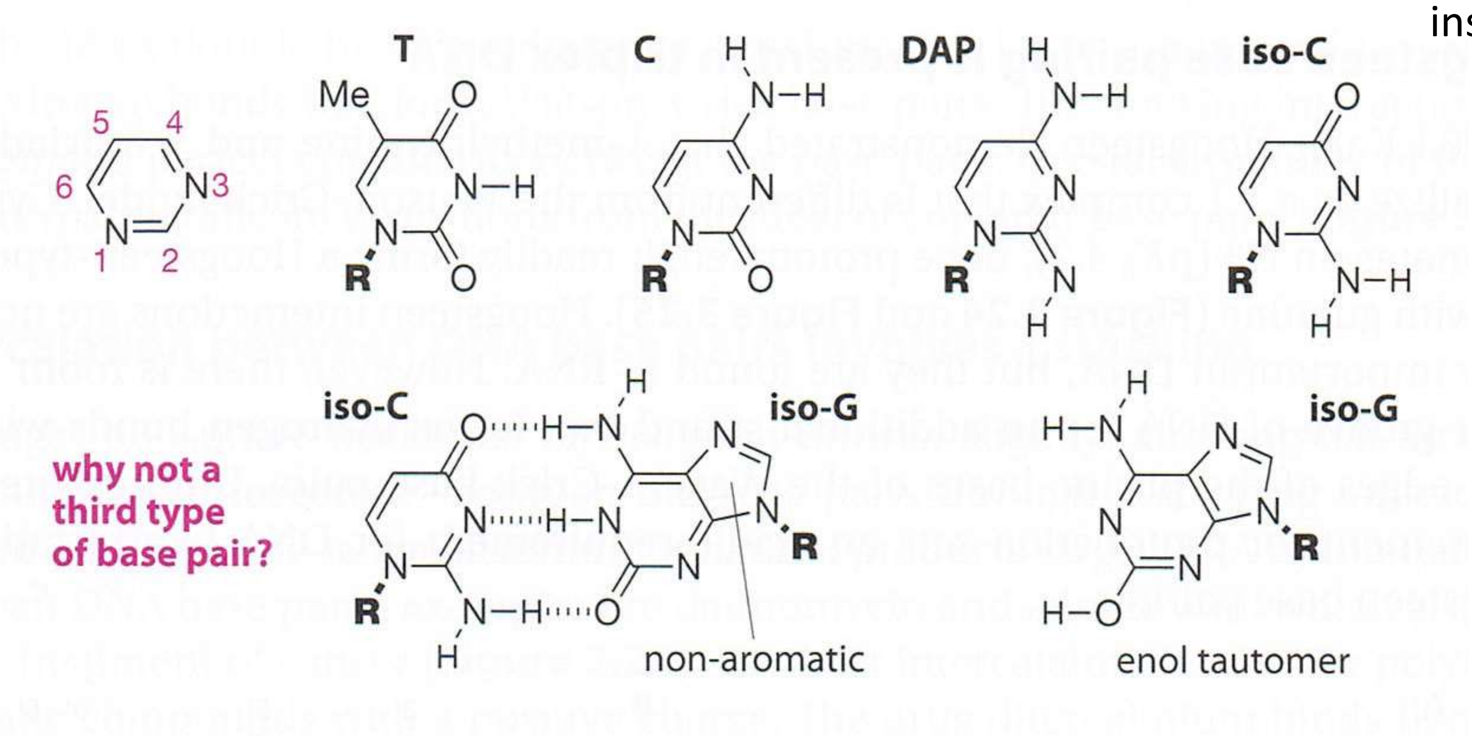
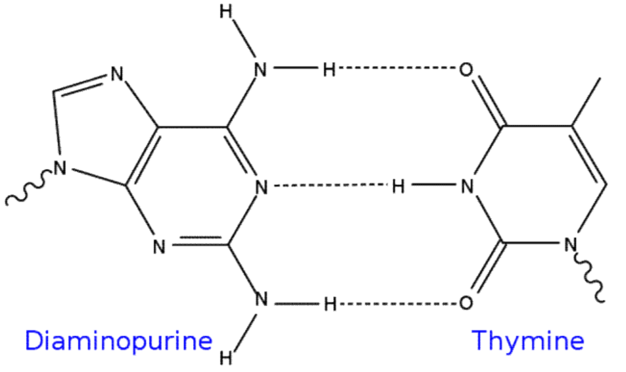


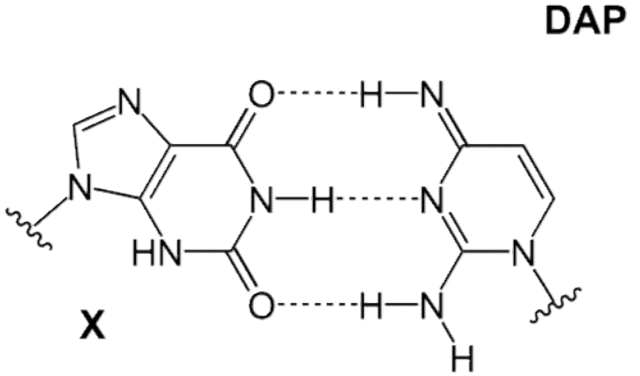
Figure 3.20 Chromosome 1 from hamster cells exposed to bromodeoxyuridine. (A) Normal chromosome. (B–E) Aberrant chromosomes. (From T.C. Hsu and C.E. Somers, *Proc. Natl. Acad. Sci. USA* 47: 396–403, 1961. With permission from the MD Anderson Cancer Center.)

Alternative base pairs – synthetic biology

Cyanophage S-2L uses diaminopurine instead of adenine (3 H-bonds!)





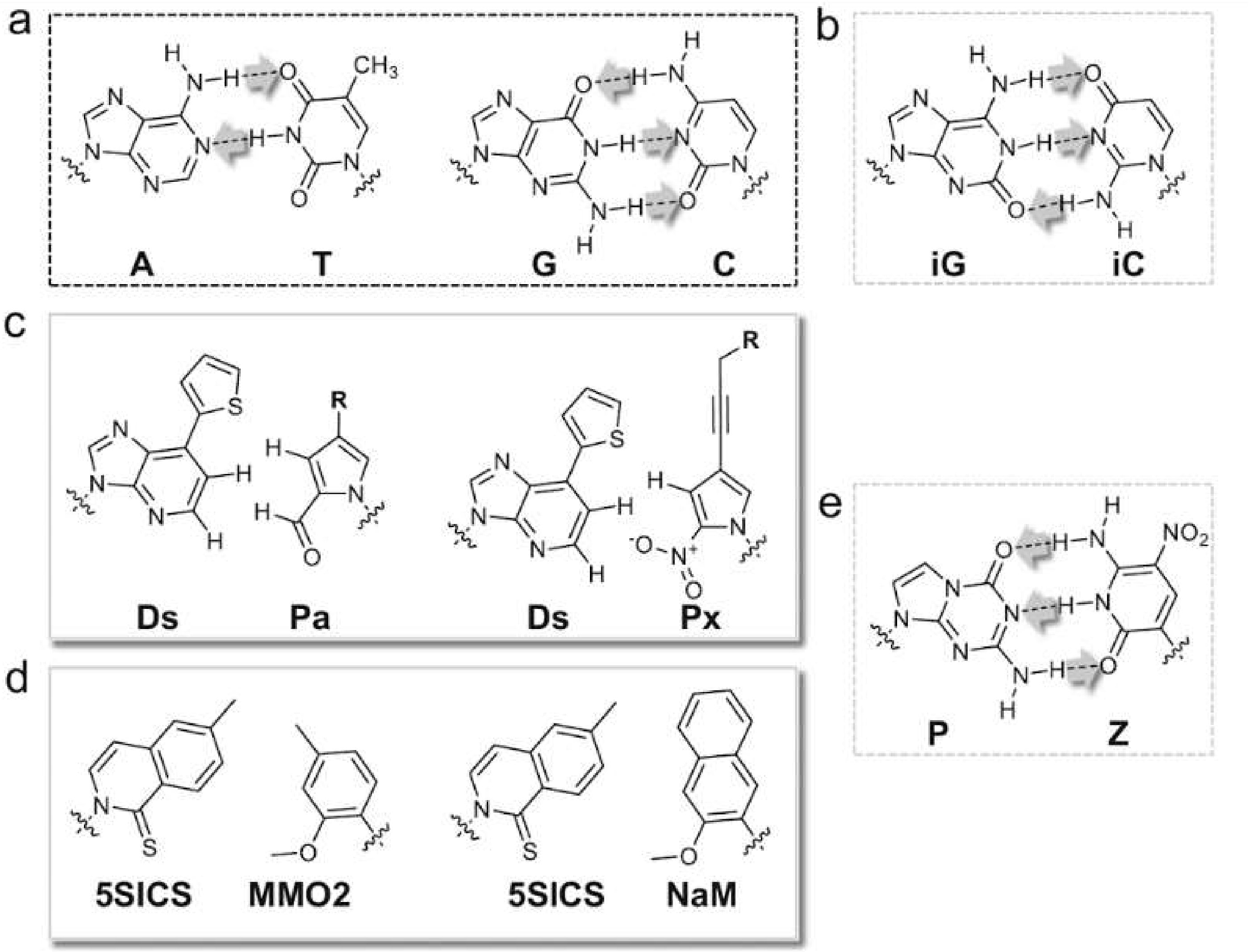


DAP – one tautomer forms a base pair with guanine

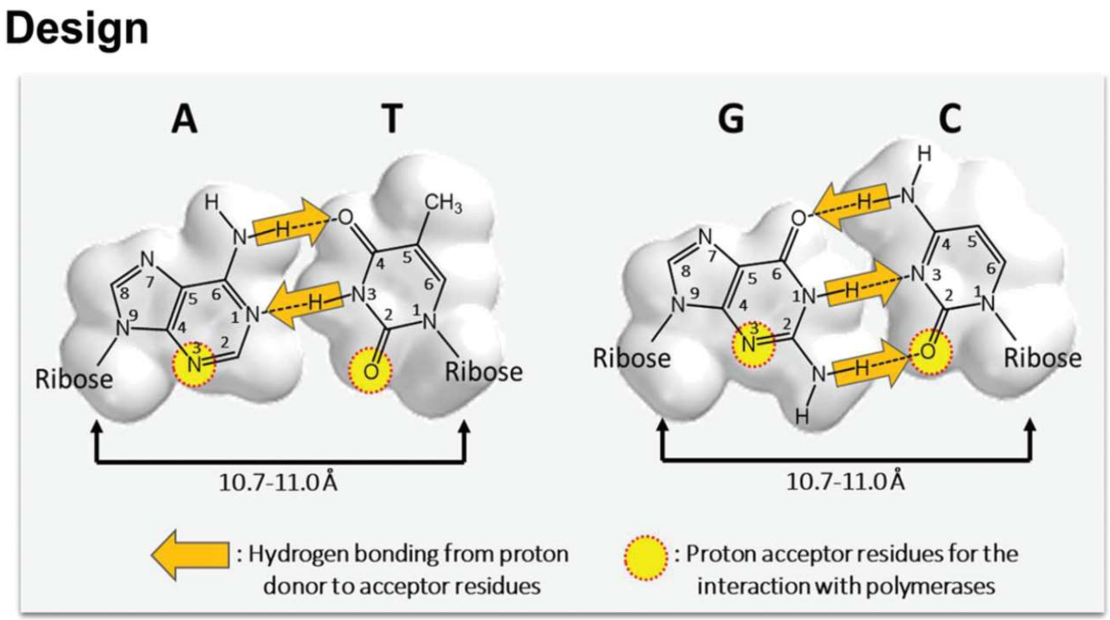
iso-C/iso-G

- specificity (the enol tautomer of iso-G, stabilized by aromatization, complementary to thymine)
- the 2-amino group of iso-C hydrolyses easily to uracil

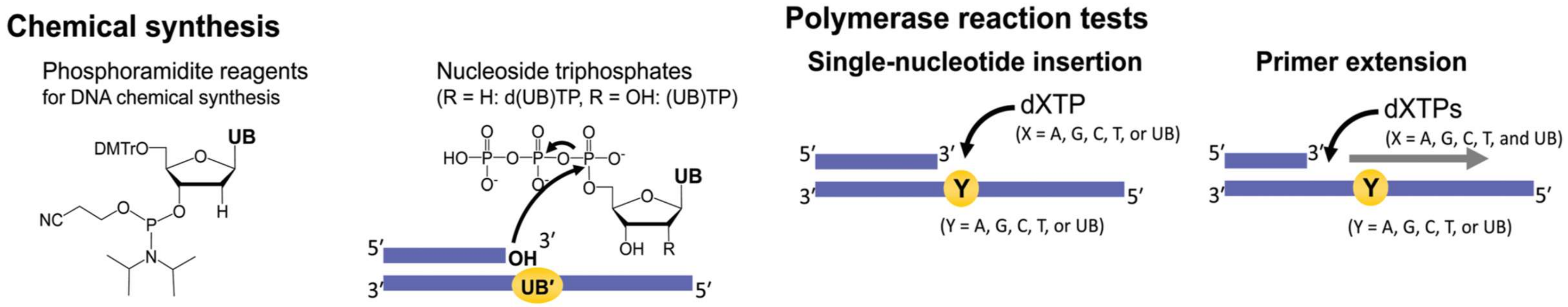
Natural and non-natural base pairs that function in polymerase reactions



Unnatural base pair (UBP) design rules:

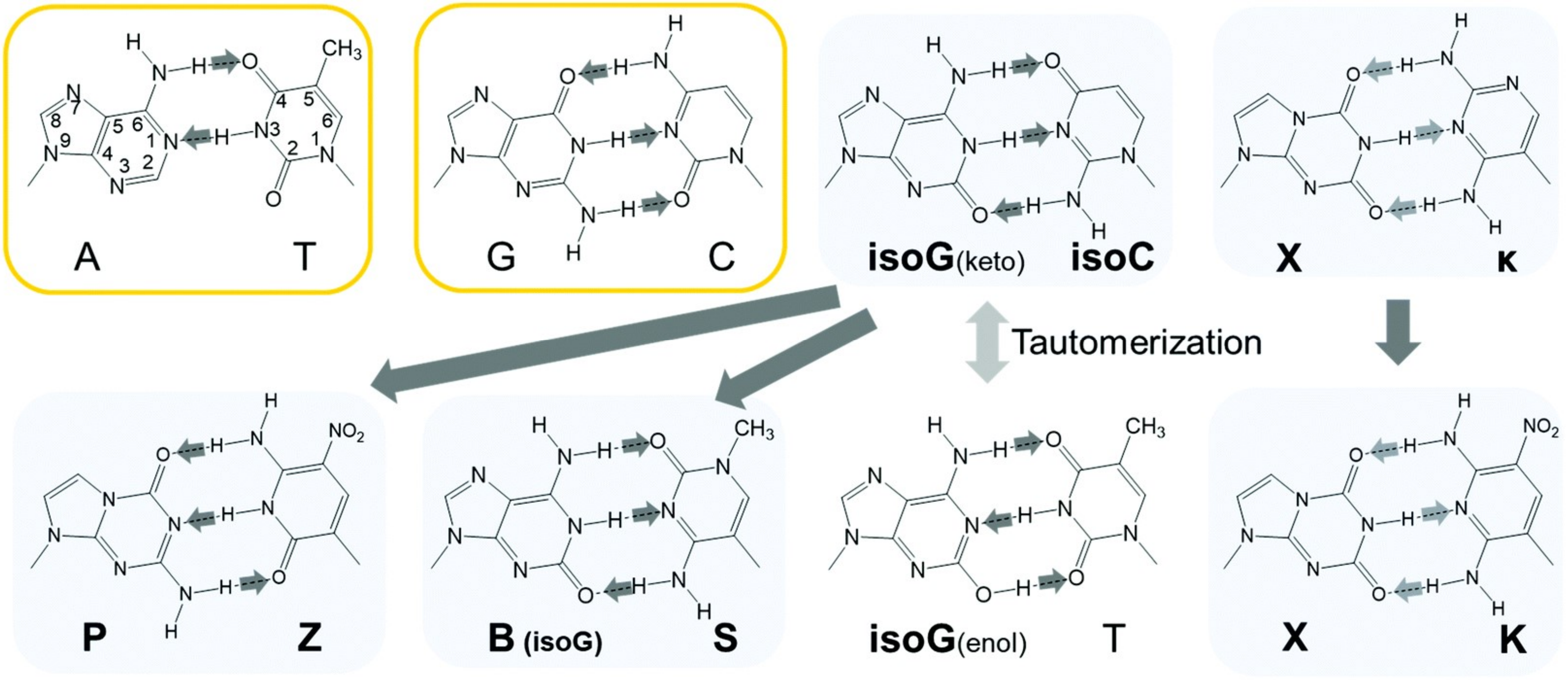


- Distance of 10.7-11.0 Å between the glycosidic bonds of the pair
- no mispairing with natural bases
- Reasonable hydrophobicity
- Chemical stability (phosphoramidite chemistry)
- Recognition by DNA and RNA polymerases (fidelity, efficiency)



M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

Non-natural nucleobases compatible with polymerases



Alexander Rich (1962): isoG–isoC pair

Steven Benner (1989-95): the artificially expanded genetic information system (AEGIS) including the isoG–isoC and X–κ pairs
→ *in vitro* replication, transcription, and translation systems.

Benner and Prudent (2004): new quantitative PCR (qPCR) methods, such as Plexor, using the isoG–isoC pair

Benner (2007): Z-P pair - the P base by removing the hydrogen at position 1 of G to exclude the keto–enol tautomerism, by introducing the nitro group into Z, the chemical stability of the nucleoside was improved;

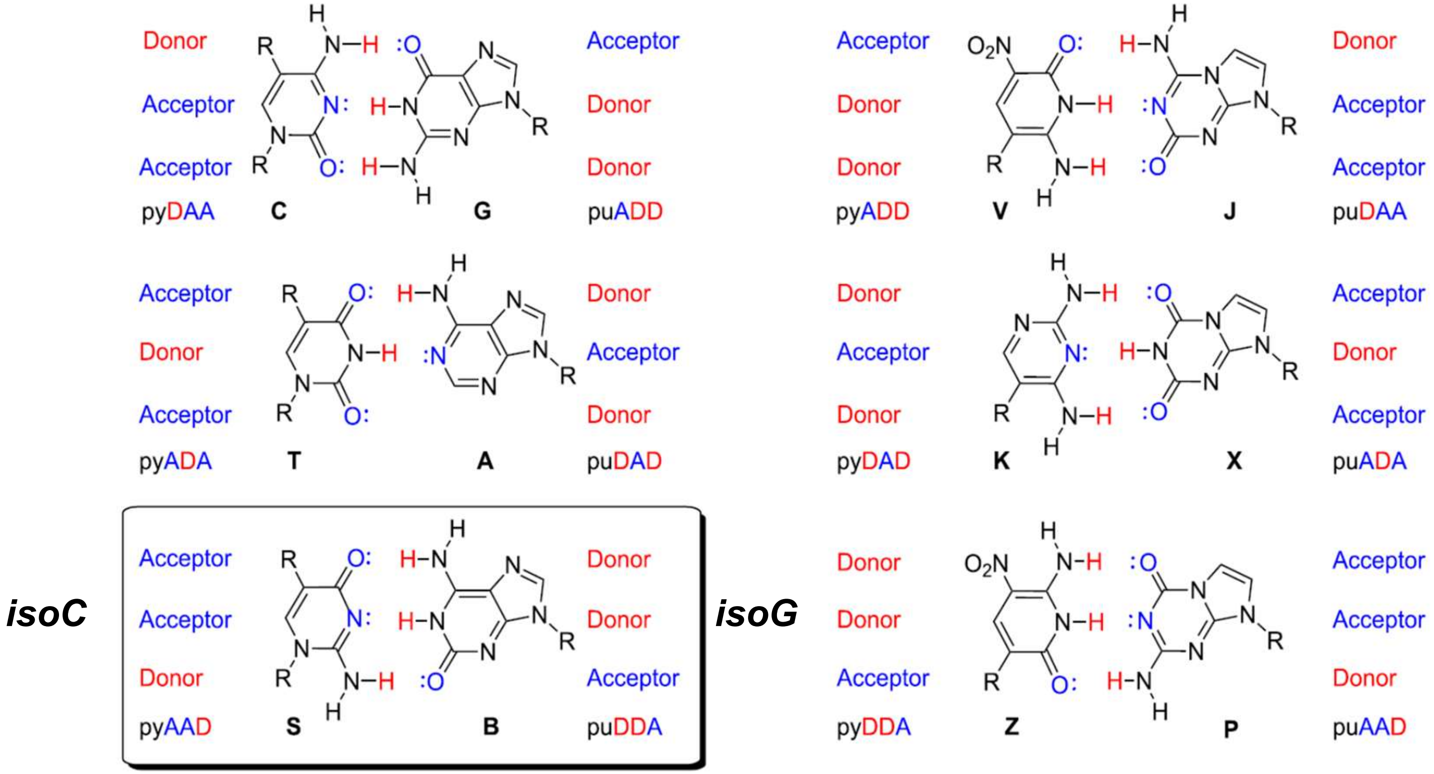
With the same strategy, they also developed the B–S and X–K pairs from isoG–isoC and X–κ pairs, respectively

AEGIS – Artificially Expanded Genetic Information System

Watson–Crick pairing rules:

(a) size complementarity - large purines pair with small pyrimidines

(b) hydrogen-bonding complementarity (hydrogen-bond acceptors, A, pair with hydrogen-bond donors, D).



Rearranging donor and acceptor groups on the nucleobases, while not changing the geometry of the Watson–Crick pair, creates an artificially expanded genetic information system (AEGIS). AEGIS components add information density to DNA strands built from them.

S. Benner *et al., Beilstein J. Org. Chem.* **2014,** *10,* 2348–2360. doi:10.3762/bjoc.10.245

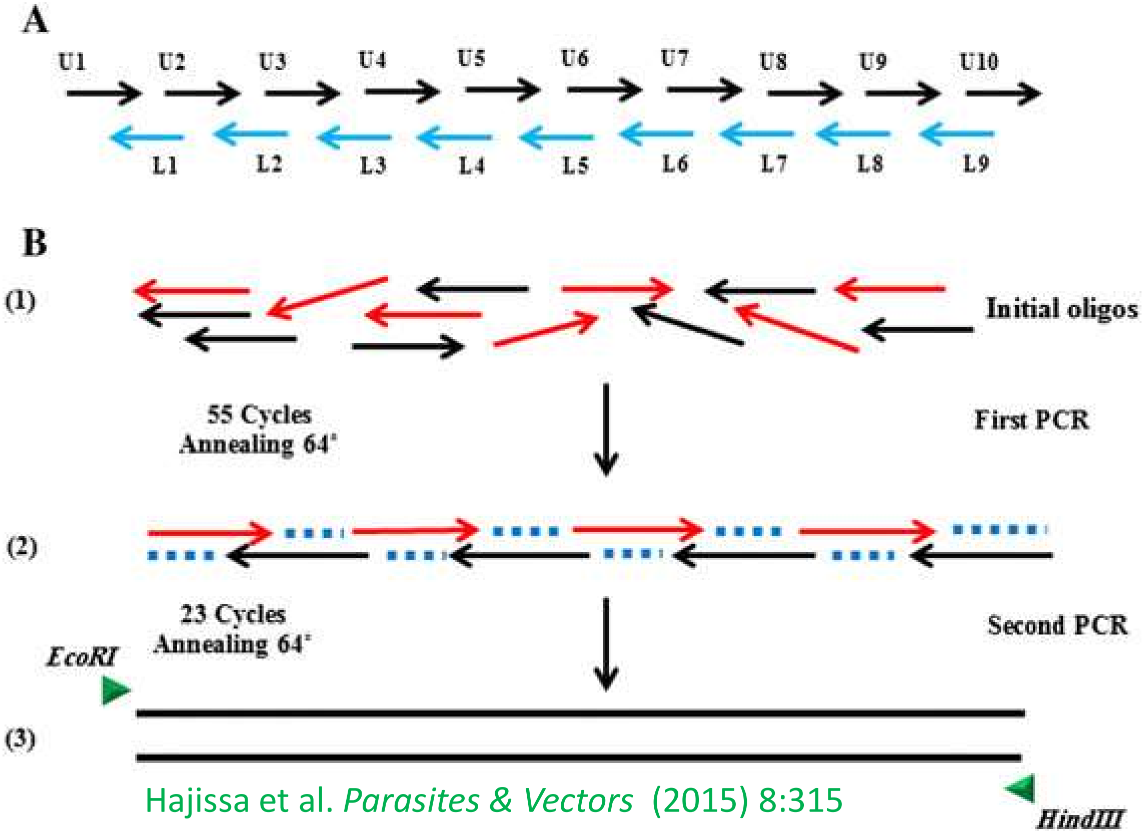
Artificial Gene Synthesis

Artificial gene synthesis (DNA printing) - method in synthetic biology to create artificial genes in the laboratory:

- *currently based on solid-phase DNA synthesis,*
- *the user does not have to begin with preexisting DNA sequences.*
- *Therefore, it is possible to make a completely synthetic double-stranded DNA molecule with no apparent limits on either nucleotide sequence or size.*

Applications:

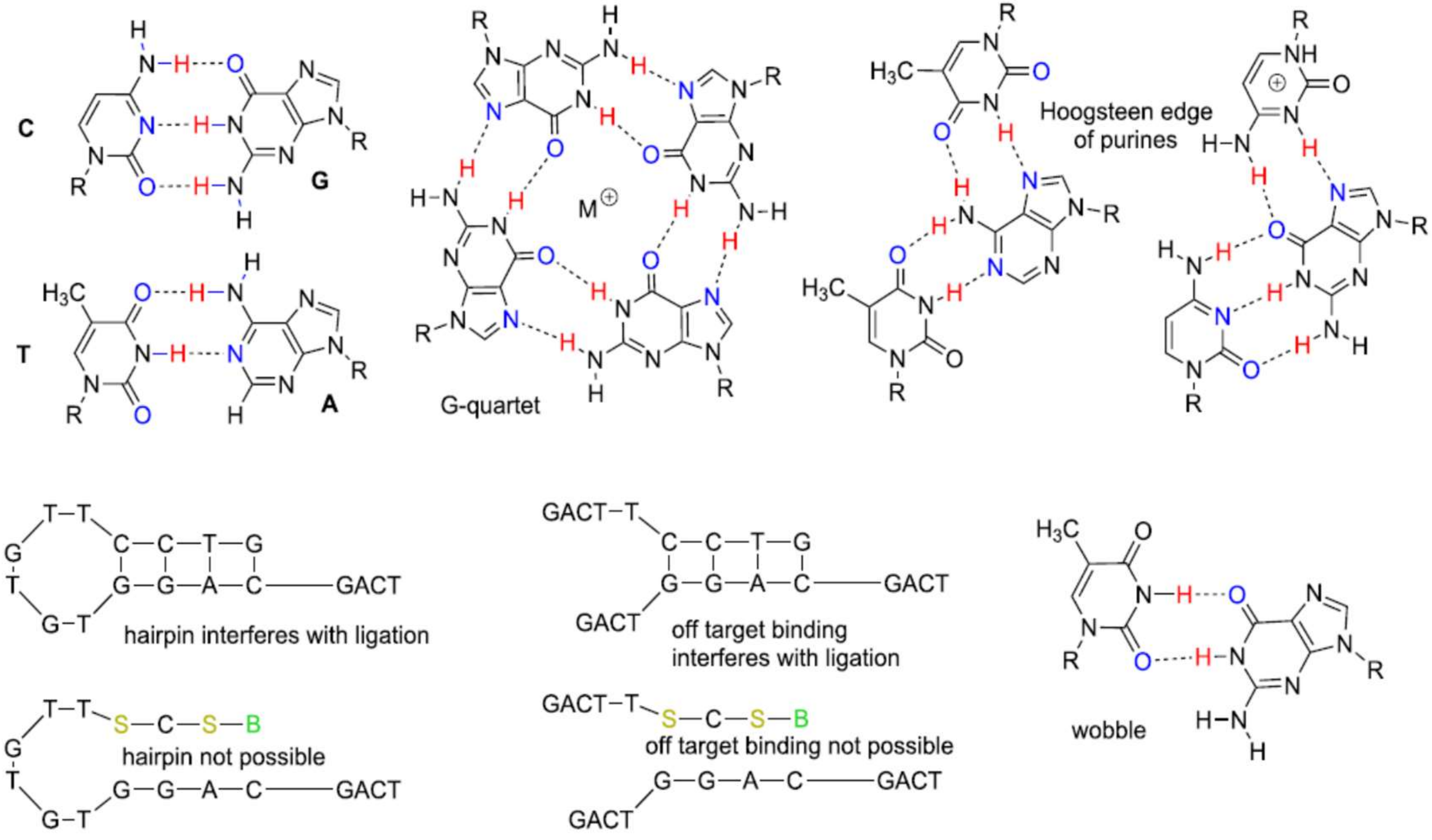
- recombinant DNA technology including heterologous gene expression, vaccine development, gene therapy and molecular engineering.
- The synthesis of nucleic acid sequences can be more economical than classical cloning and mutagenesis procedures
- the ability to safely obtain genes for vaccine research without the need to grow the full pathogens.
- to optimize protein expression in a particular host, or to remove non-functional DNA segments
- For DNA digital data storage and computing
- For synthetic biological circuits



Hajissa et al. *Parasites & Vectors* (2015) 8:315

Self-assembly of whole genes and DNA nanostructures

Limitations of DNA puzzle assembly: unequal A:T vs. G:C strength, insufficient ACGT information density, higher-order structures

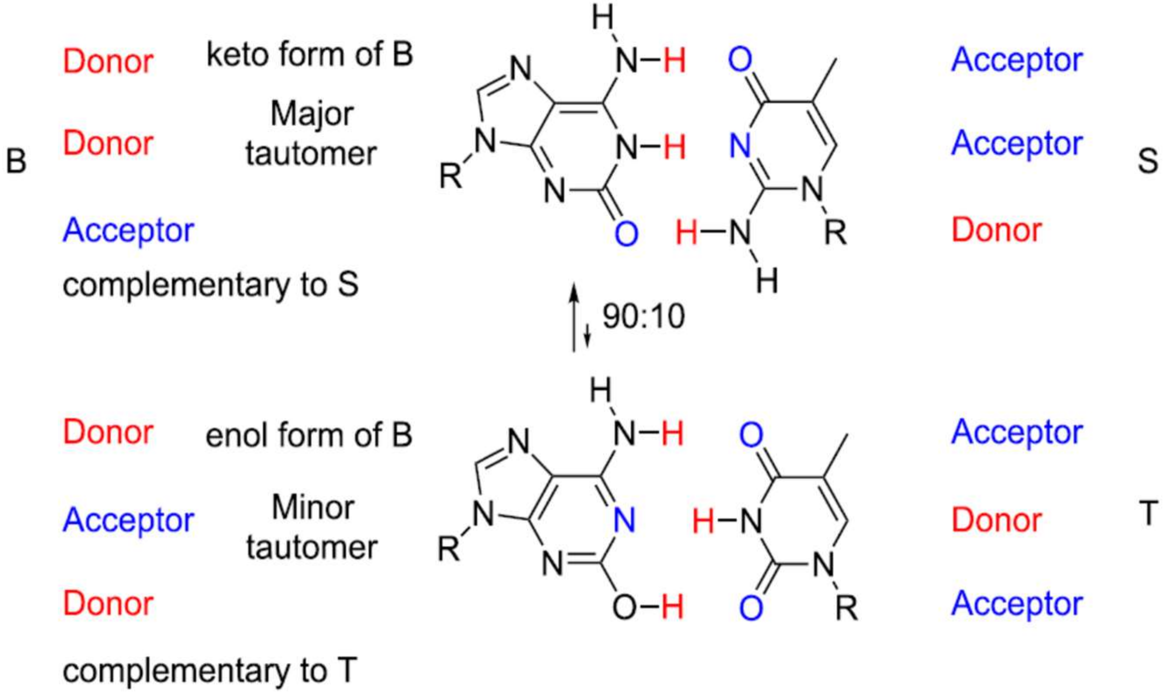


S. Benner *et al., Beilstein J. Org. Chem.* **2014,** *10,* 2348–2360. doi:10.3762/bjoc.10.245

Self-assembly of whole genes and DNA nanostructures

Solution: an orthogonal pair from the AEGIS system, that can be removed from the product, yielding native DNA structures

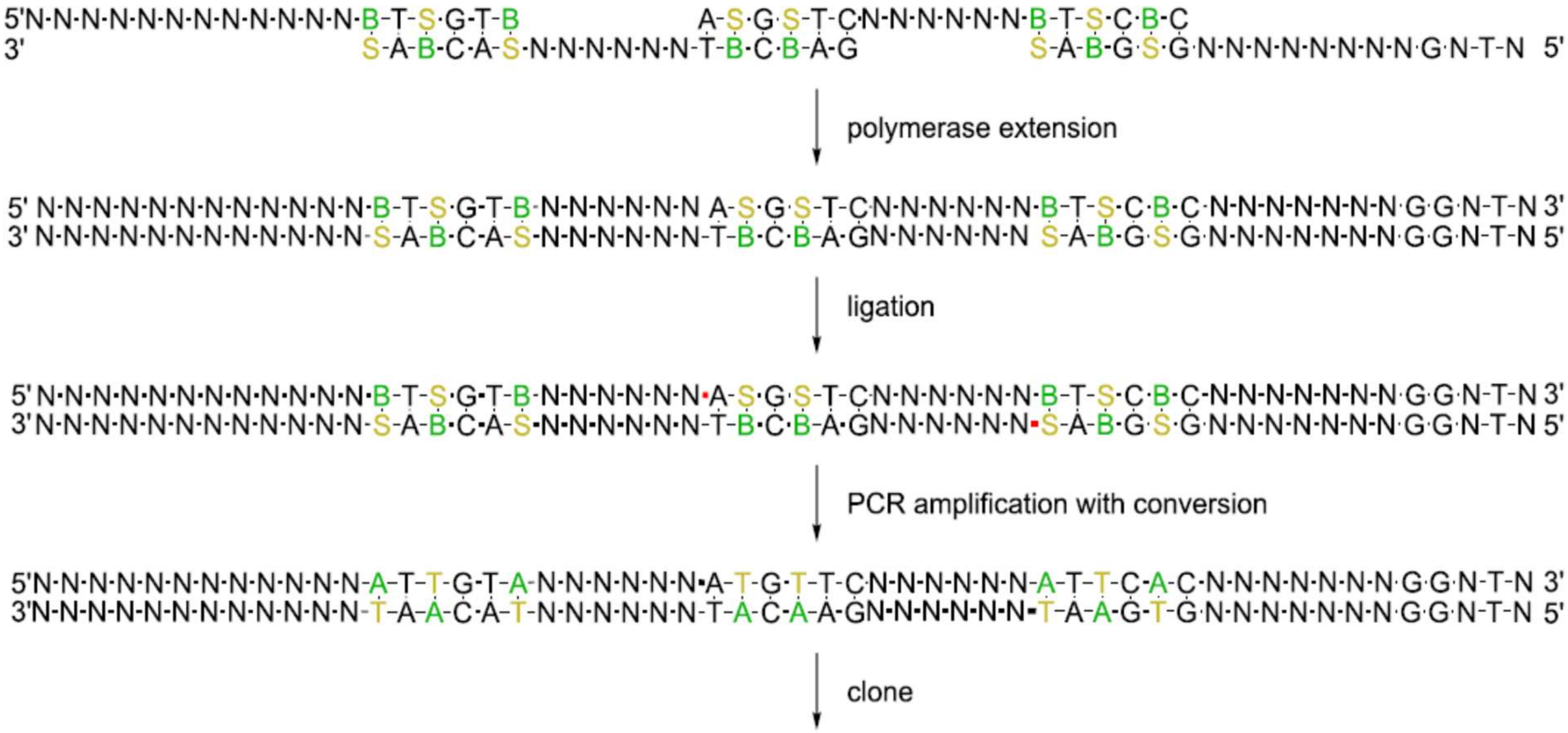
Conversion occurs when polymerases are forced to mismatch a standard nucleotide opposite an AEGIS nucleotide by
(a) not being provided the complementary AEGIS triphosphate and
(b) exploiting a chemical feature of the AEGIS nucleotide that directs a specific mismatch.



B in its major tautomeric form pairs with **S**; in its minor tautomeric form, **B** pairs with standard **T**. Asembly of the target gene/DNA nanostructure is followed by conversion of the **S:B** pairs to **T:A** pairs after two cycles of PCR:
B → **A** via an intermediate **B:T** mispairing, **S** → **T** (intermediate **S:B** followed by a second **B:T** mispairing).

S. Benner *et al., Beilstein J. Org. Chem.* **2014,** *10,* 2348–2360. doi:10.3762/bjoc.10.245

Self-assembly of whole genes and DNA nanostructures

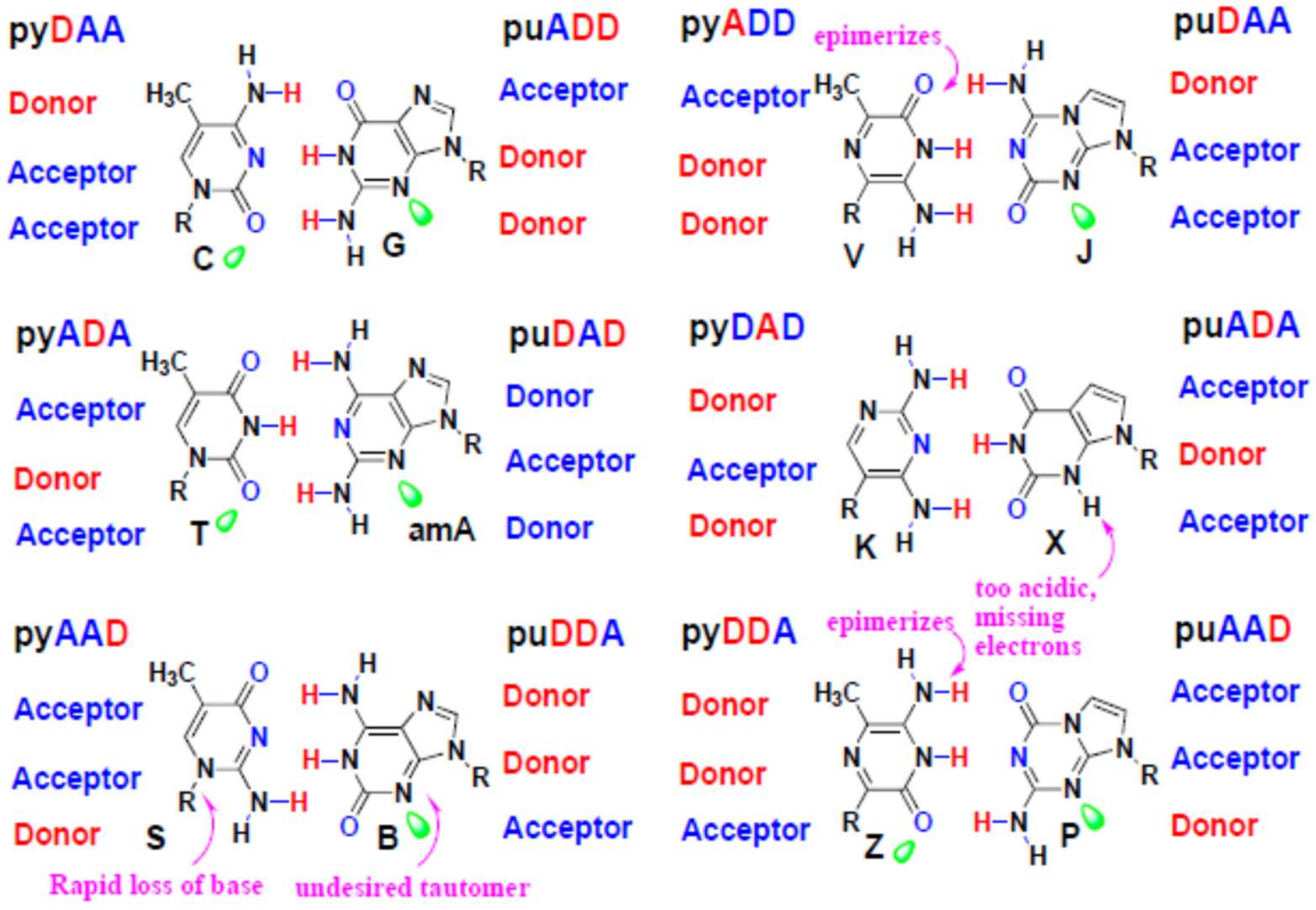


The technology tested by assembly of the kanamycin-resistance gene and growing the bacteria in the environment containing kanamycin after assembly and conversion of that gene.

S. Benner *et al., Beilstein J. Org. Chem.* **2014,** *10,* 2348–2360. doi:10.3762/bjoc.10.245

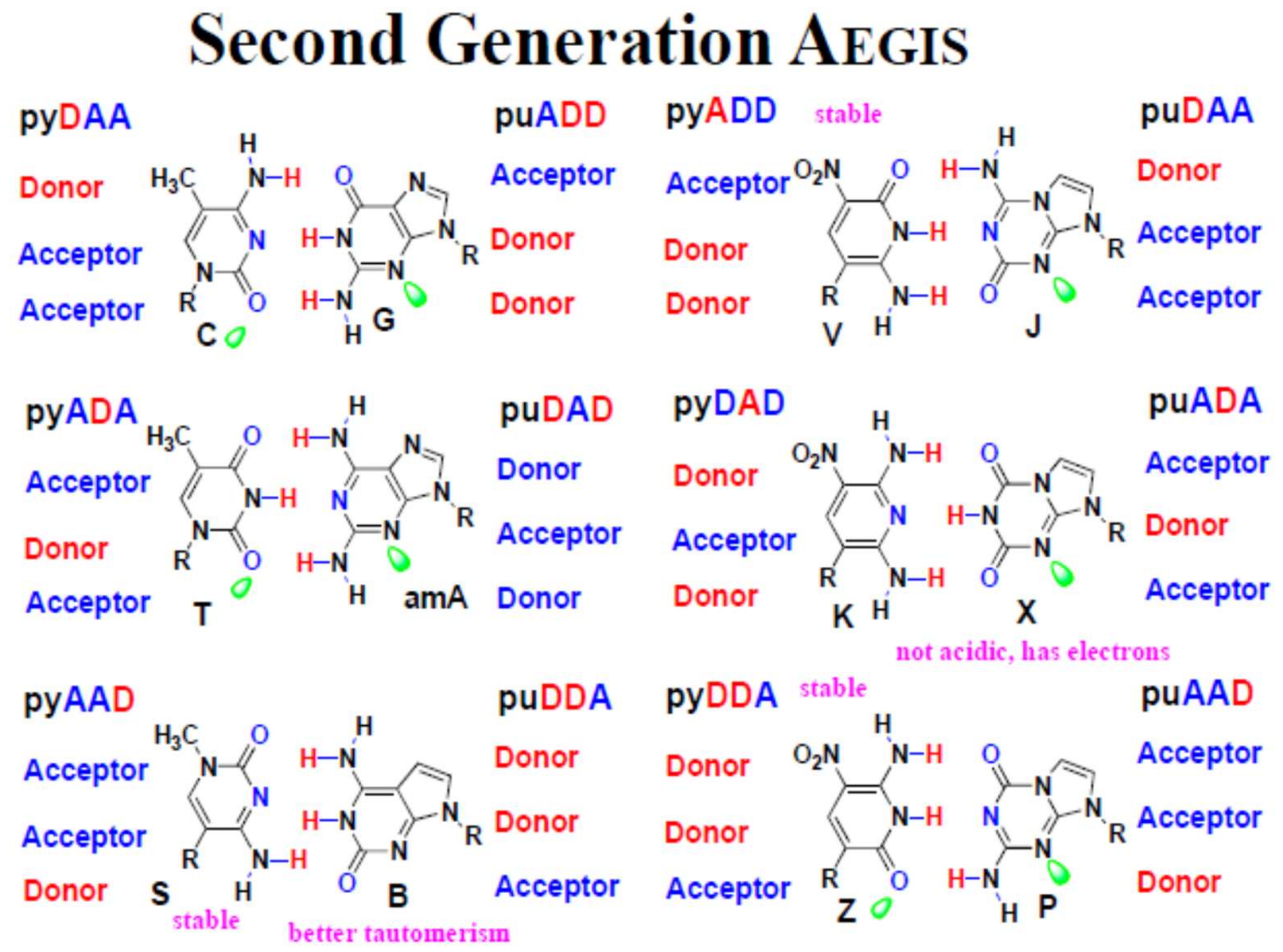
AEGIS – Artificially Expanded Genetic Information System

First Generation AEGIS



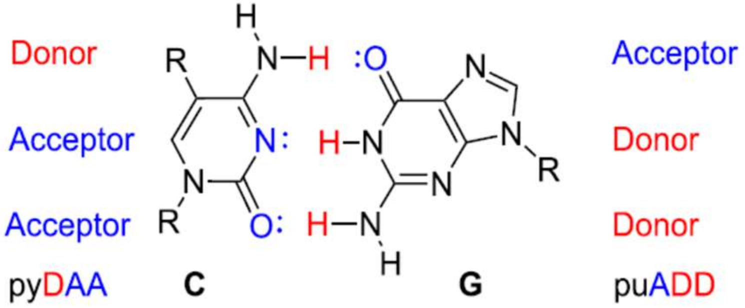
E. Biondi, S. Benner *Biomedicines* **2018**, *6*, 53; *doi:10.3390/biomedicines6020053*

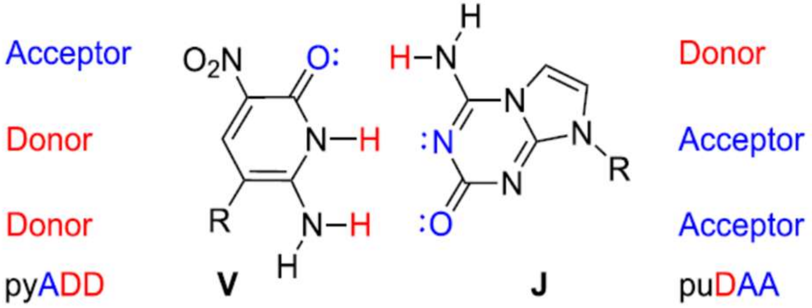
AEGIS – Artificially Expanded Genetic Information System

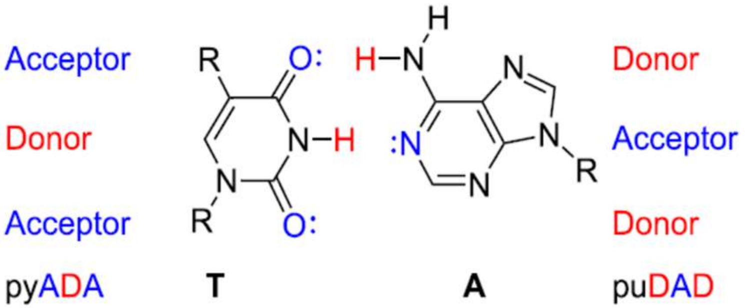


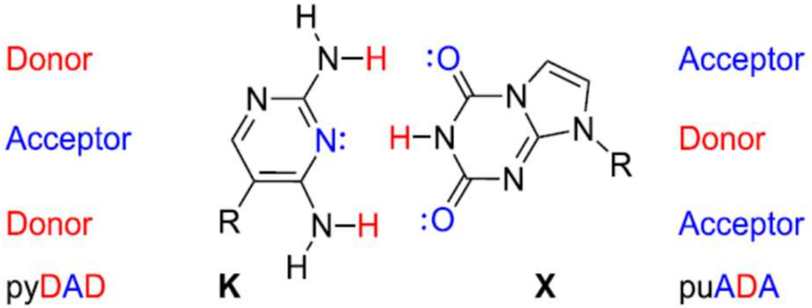
E. Biondi, S. Benner *Biomedicines* **2018**, *6*, 53; *doi:10.3390/biomedicines6020053*

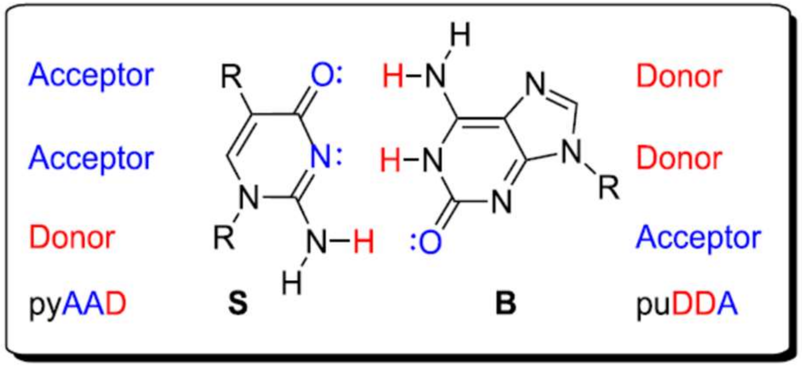
AEGIS – Artificially Expanded Genetic Information System

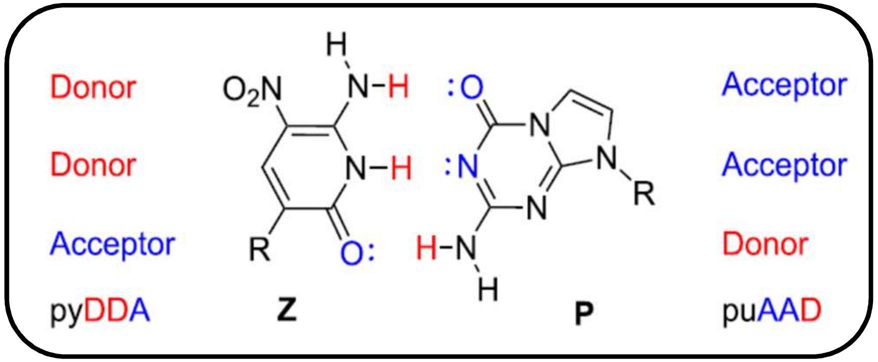






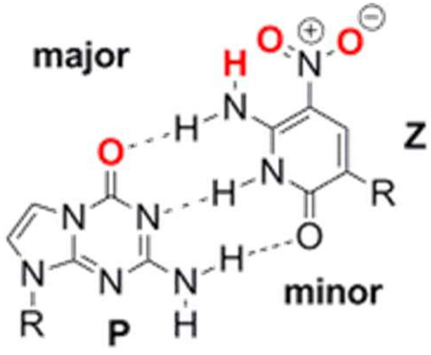


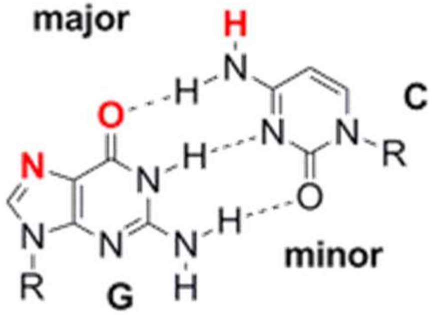


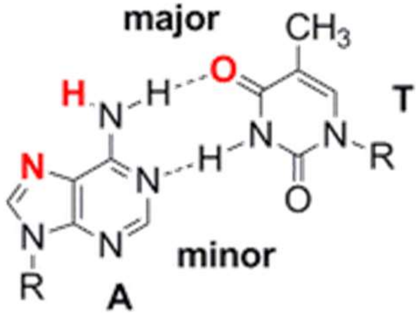


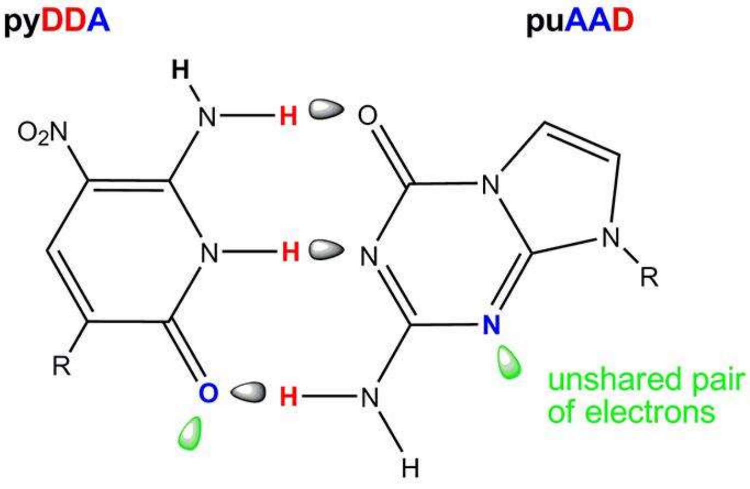
S. Benner *et al., Beilstein J. Org. Chem.* **2014,** *10,* 2348–2360. doi:10.3762/bjoc.10.245

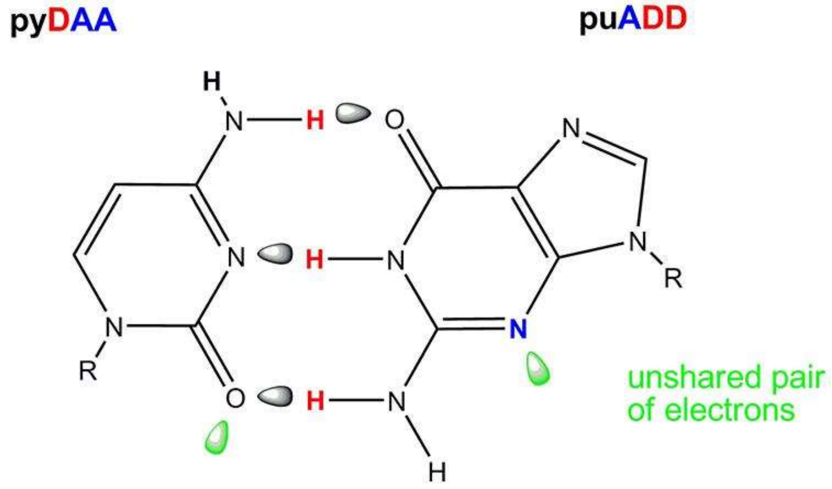
AEGIS – Artificially Expanded Genetic Information System







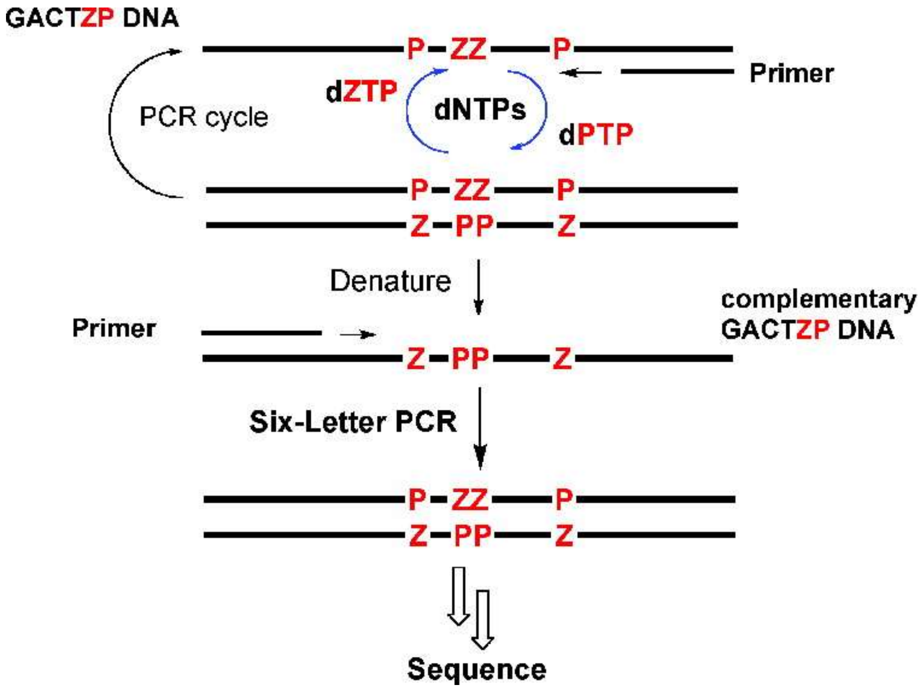




Z **P** **C** **G**

S. Benner *et al., J. Am. Chem. Soc.,* **2011**, *133 (38),* pp 15105–15112

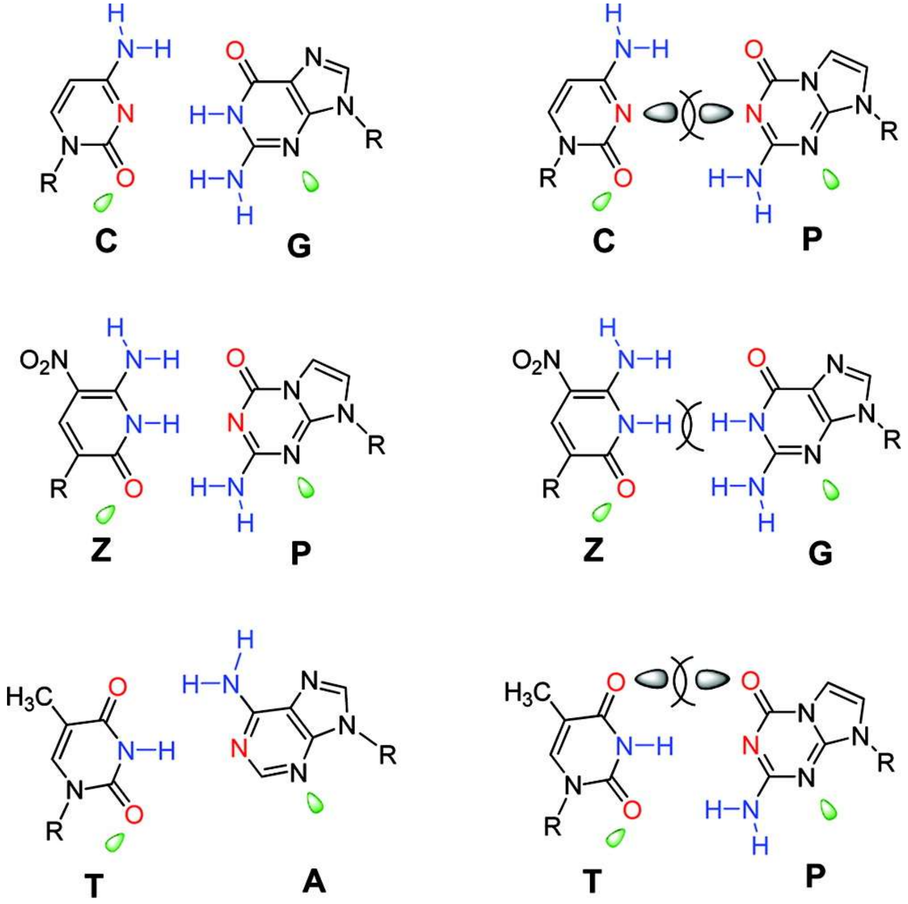
AEGIS – Permanent orthogonal nucleobases surviving PCR



Error rate 0,2% per a PCR cycle – both removal and incorporation of **Z** and **P** ➔ the artificial genetic system capable to evolve.

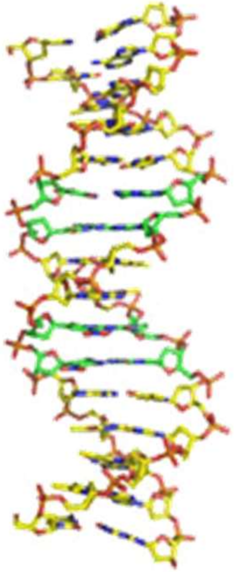
Pol: Deep Vent – 2 Z/P, Taq/Phu – 3-4 Z/P
dZTP (*deprotonated*) at higher pH pairs slightly with G
→ loss of some **Z**, but also gain of some new **Z** mutants.

Electron density presented to the minor groove
→ recognition site by polymerases
„ minor groove scanning hypothesis"

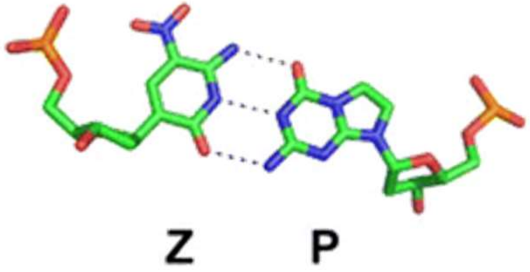


S. Benner *et al., J. Am. Chem. Soc.*, **2011**, *133 (38),* pp 15105–15112

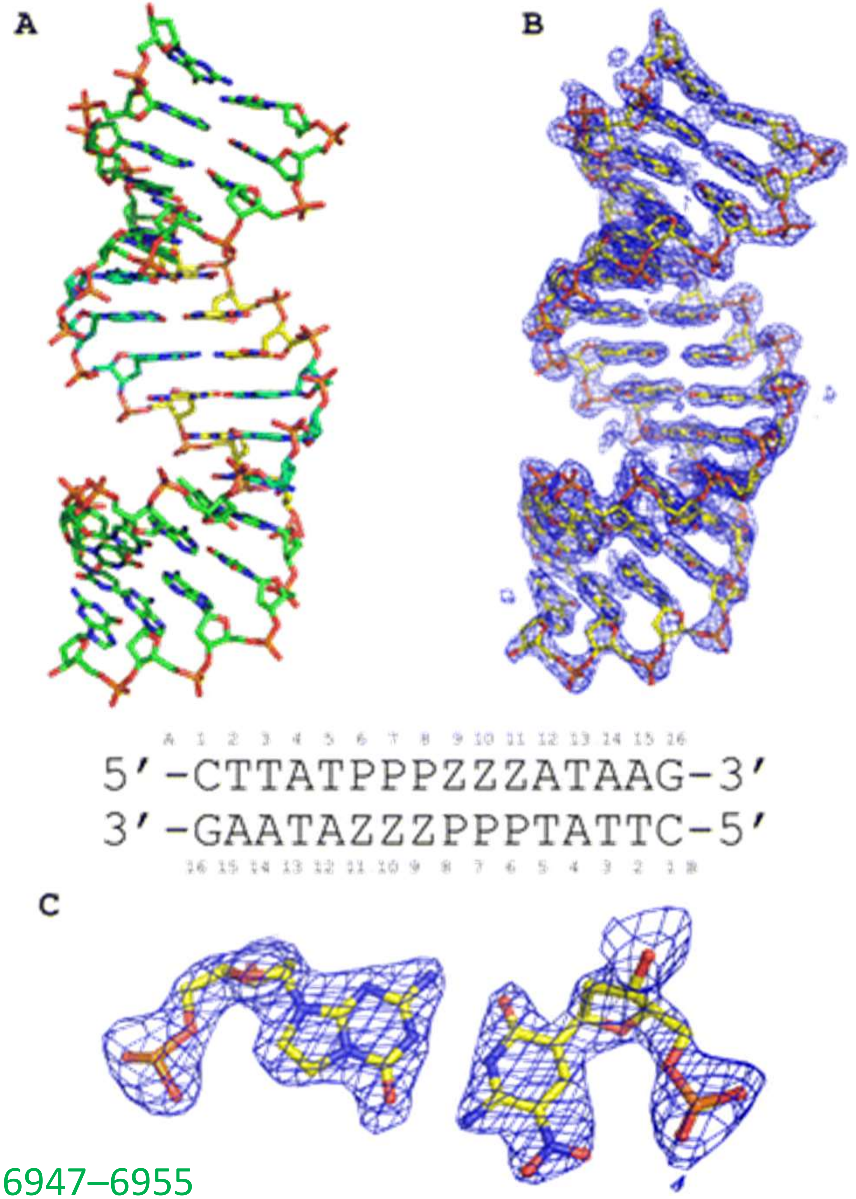
ACGTZP-DNA crystal structures



B-DNA



A-DNA

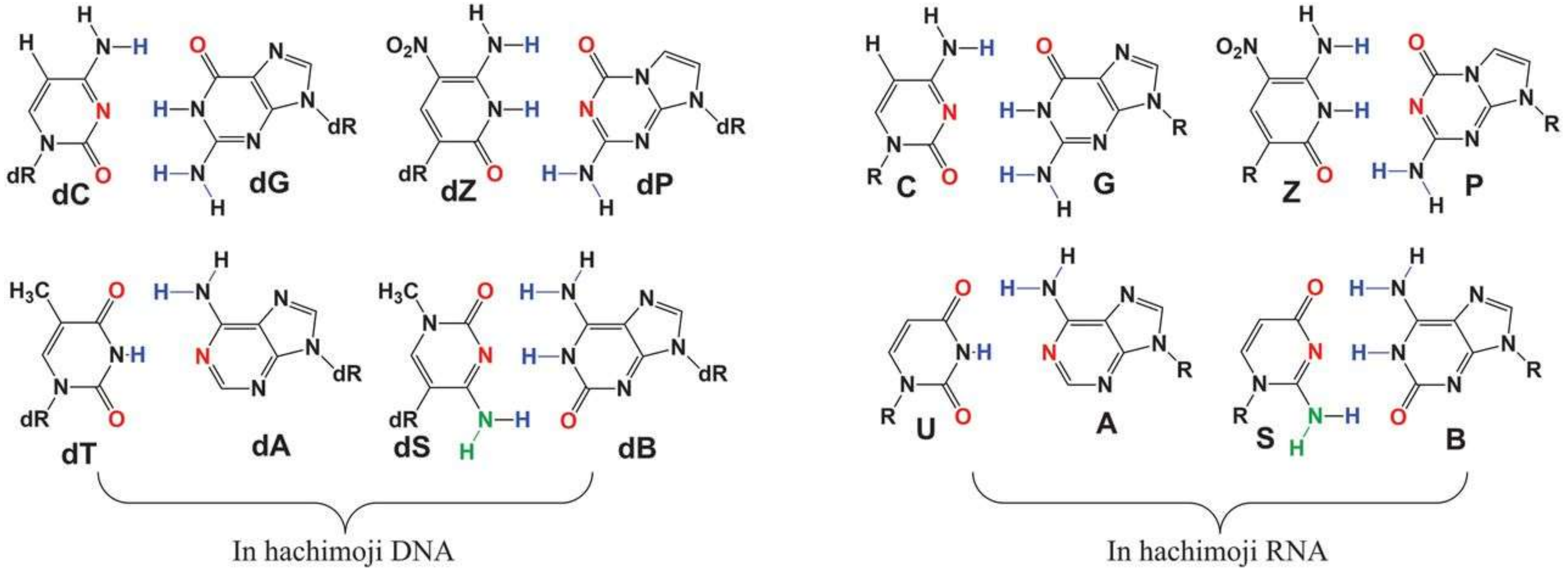


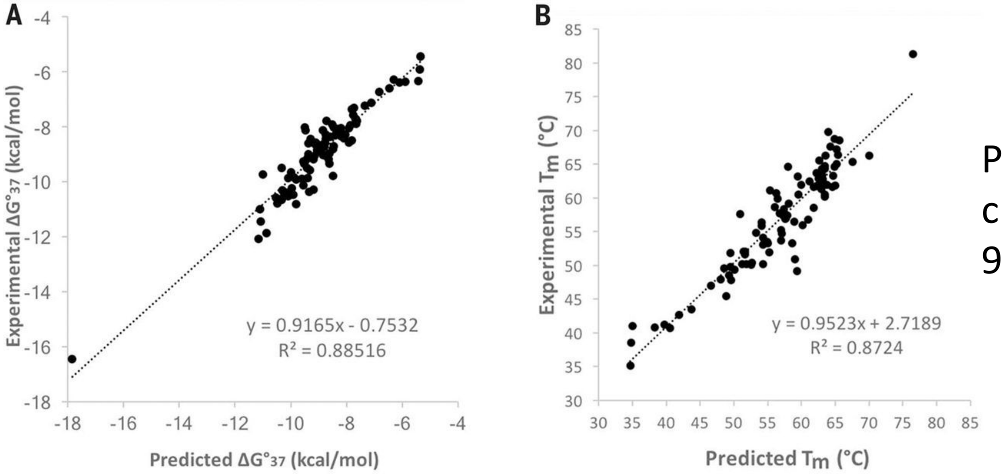
5′-CTTATPPPZZZATAAG-3′
3′-GAATAZZZPPPTATTC-5′

*18-mers: 2+2 **Z:P** pairs → B-DNA*
*6 consecutive **Z:P** → A-DNA*
*0,1 nm wider, but otherwise alike **G:C** pairs*

S. Benner *et al., J. Am. Chem. Soc.,* **2015**, *137,* pp 6947–6955

Hachimoji DNA and RNA – a genetic system with eight (Jap.- Hachi) letters



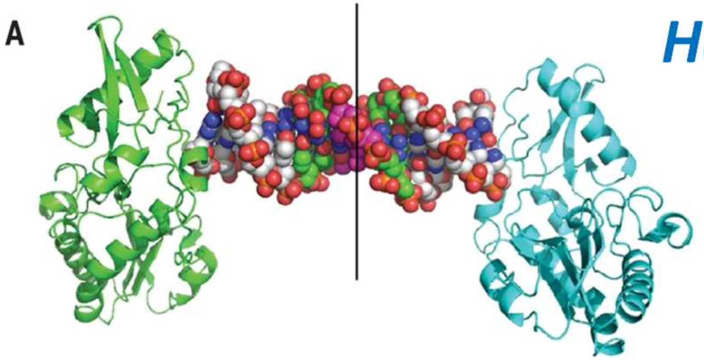


Plots of experimental versus predicted **(A)** free energy changes (ΔG°_{37}) and **(B)** melting temperatures T_m for 94 SBZP-containing hachimoji DNA duplexes.

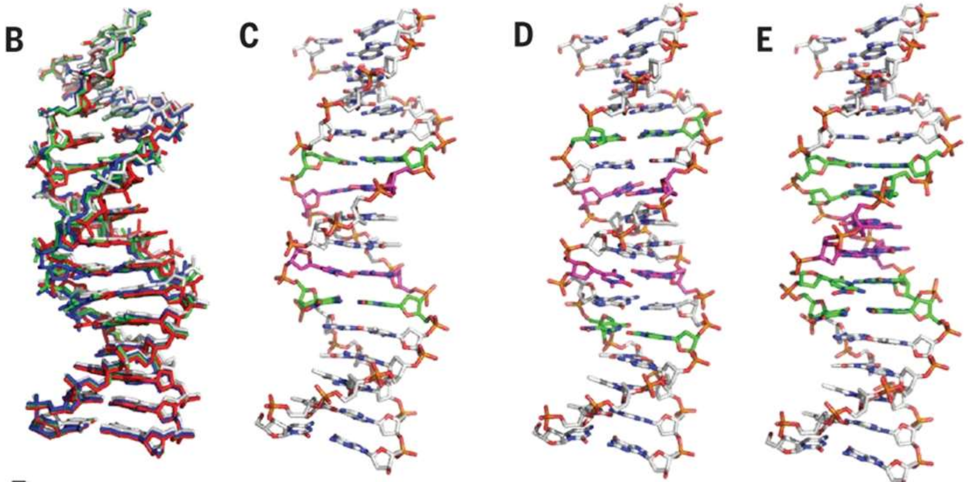
S. Hoshika *et al.*, *Science* **2019**, *363*, 884-887

Hachimoji DNA and RNA

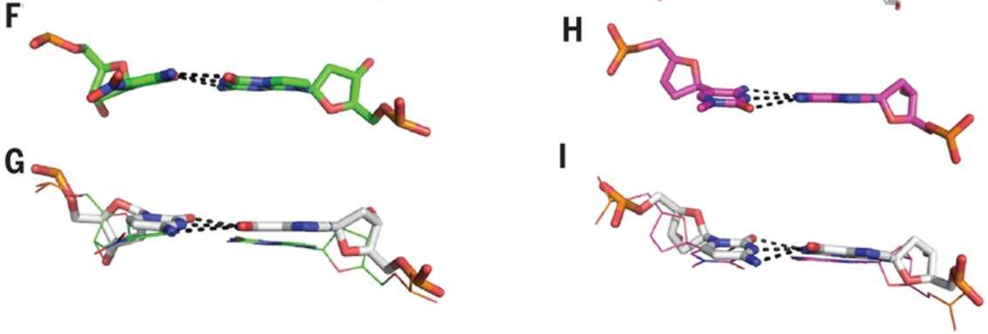
Crystal structures of hachimoji DNA.



(**A**) The host-guest complex with two N-terminal fragments from Moloney murine leukemia virus reverse transcriptase bound to a 16-mer PP hachimoji DNA; Z:P pairs are green and S:B pairs are magenta.



(**B**) Hachimoji DNA structures PB (green), PC (red), and PP (blue) are superimposed with GC DNA (gray).
(**C**) Structure of hachimoji DNA with self-complementary duplex 5′-CTTATPBTASZATAAG (“PB”).
(**D**) Structure of hachimoji DNA with self-complementary duplex 5′-CTTAPCBTASGZTAAG (“PC”).
(**E**) Structure of hachimoji DNA with self-complementary duplex with six consecutive nonstandard 5′-CTTATPPSBZZATAAG (PP) components.

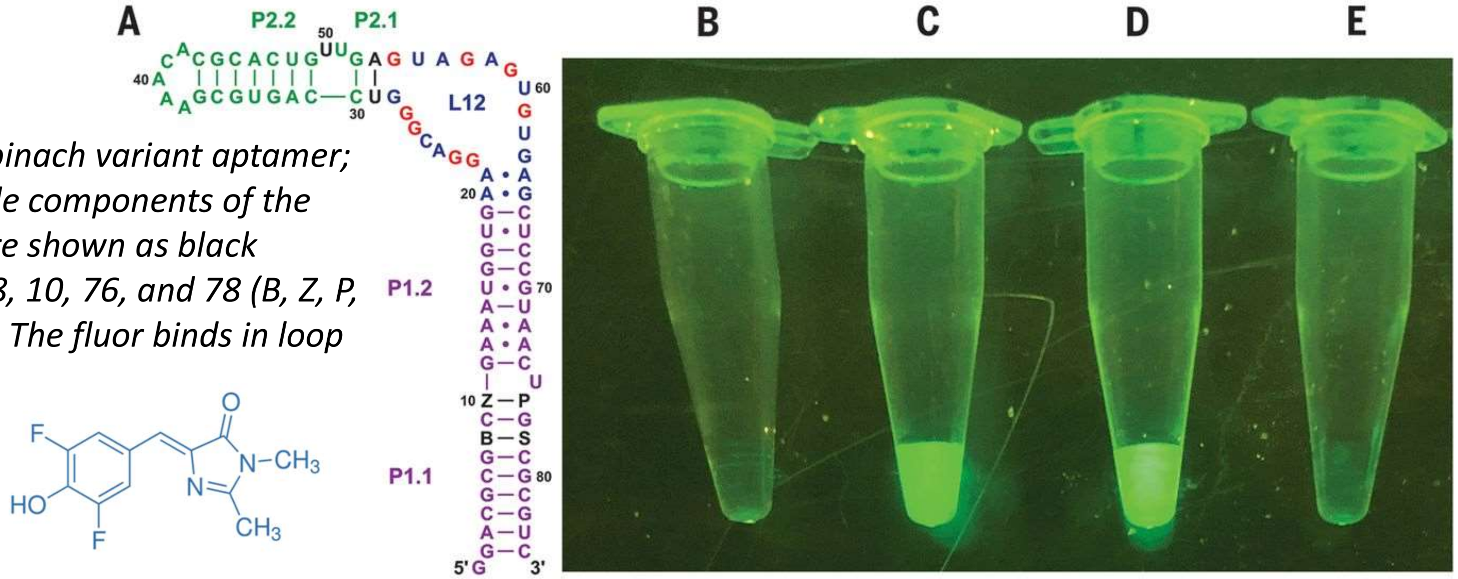


(**F** to **I**) Examples of largest differences in detailed structures. The Z:P pair from the PB structure (F) is more buckled than the corresponding G:C pair (G). The S:B pair from the PB structure (H) exhibits a propeller angle similar to that in the corresponding G:C pair (I).

S. Hoshika *et al., Science* **2019**, *363,* 884-887

Hachimoji RNA aptamer

T7 RNA polymerase incorporates ZTP, PTP, and BTP, but not STP opposite to dP, dZ, dS, and dB, respectively. A mutant of T7 RNA Pol (Y639F H784A P266L, “**FAL**”) incorporated also STP – full DNA→RNA conversion possible



the full hachimoji spinach variant aptamer; additional nucleotide components of the hachimoji system are shown as black letters at positions 8, 10, 76, and 78 (B, Z, P, and S, respectively). The fluor binds in loop L12

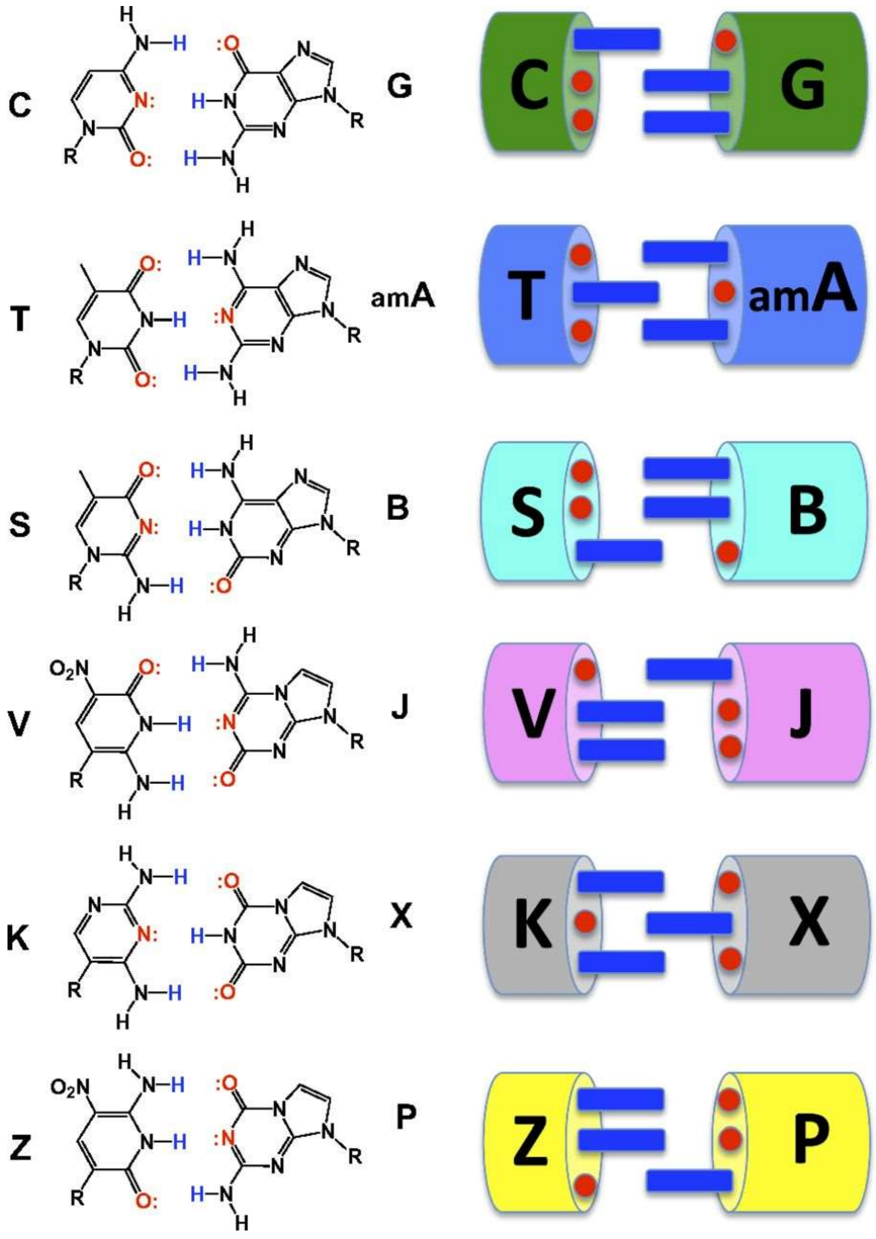
The hachimoji variant of the spinach fluorescent RNA aptamer. In its standard form, spinach folds and binds 3,5-difluoro-4-hydroxybenzylidene imidazolinone, which fluoresces green when bound.

(B) Control with fluor only, lacking RNA.
(C) Hachimoji spinach with the sequence shown in (A).
(D) Native spinach aptamer with fluor.
(E) Fluor and spinach aptamer containing Z at position 50, replacing the A:U pair at positions 53:29 with G:C to restore the triple observed in the crystal structure. This places the quenching Z chromophore near the fluor;

S. Hoshika *et al.*, *Science* **2019**, *363*, 884-887

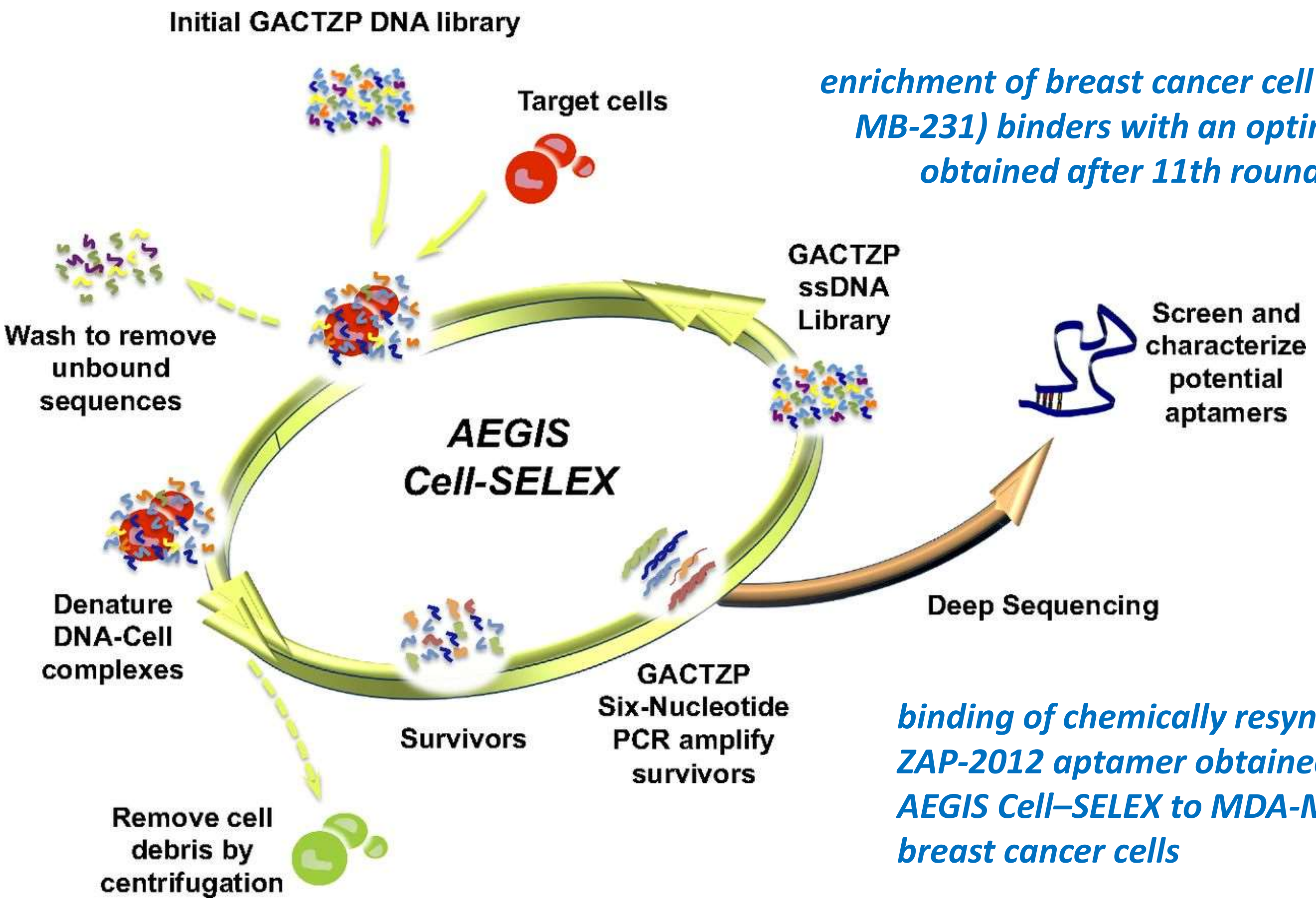
AEGIS – Artificially Expanded Genetic Information System

*An **xNA** biopolymer having functionalized AEGIS components may allow SELEX to yield protein-like aptamers better than the standard DNA and RNA biopolymers.*

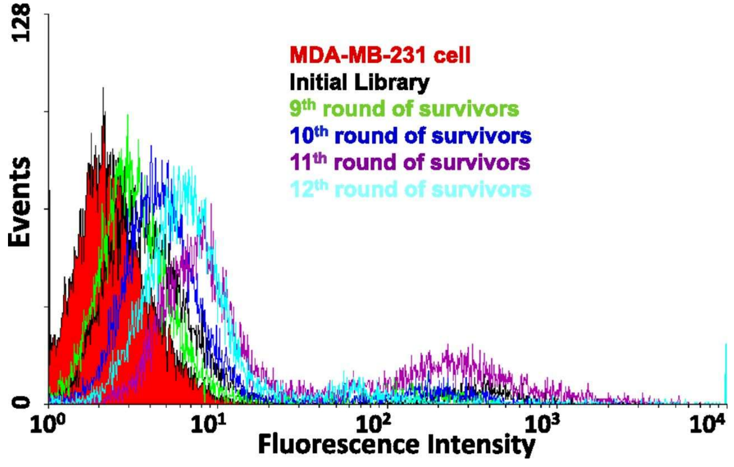


K. Sefah *et al., Proc. Natl. Acad. Sci. USA* **2014,** *111 (4),* 1449-1454.

ACGTZP-aptamers



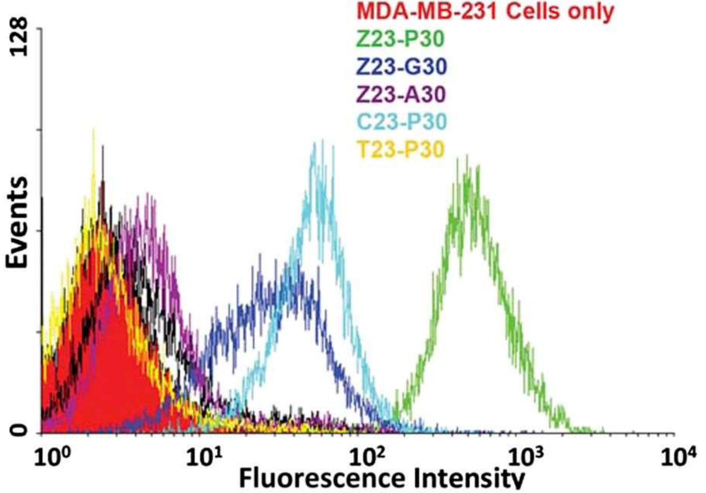
enrichment of breast cancer cell (MDA-MB-231) binders with an optimum obtained after 11th round



ZAP-2012 (Z23-P30) K_d = 30 nM

5'-biotin-TCCCGAGTGACGCAGC-
***CCCCGGZGGGATTPATCGGT**-*
GGACACGGTGGCTGAC-3'

binding of chemically resynthesized ZAP-2012 aptamer obtained using AEGIS Cell–SELEX to MDA-MB-231 breast cancer cells

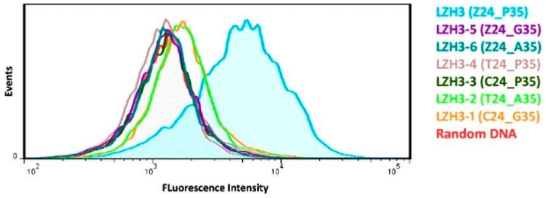


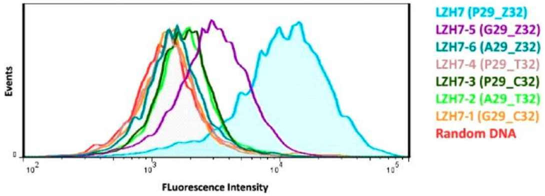
Where the **Z** or **P** are separately replaced with standard nucleotides, binding affinity is reduced. Where the **Z** and **P** are both replaced with standard nucleotides, binding affinity is lost (not shown).

K. Sefah *et al., Proc. Natl. Acad. Sci. USA* **2014,** *111 (4),* 1449-1454.

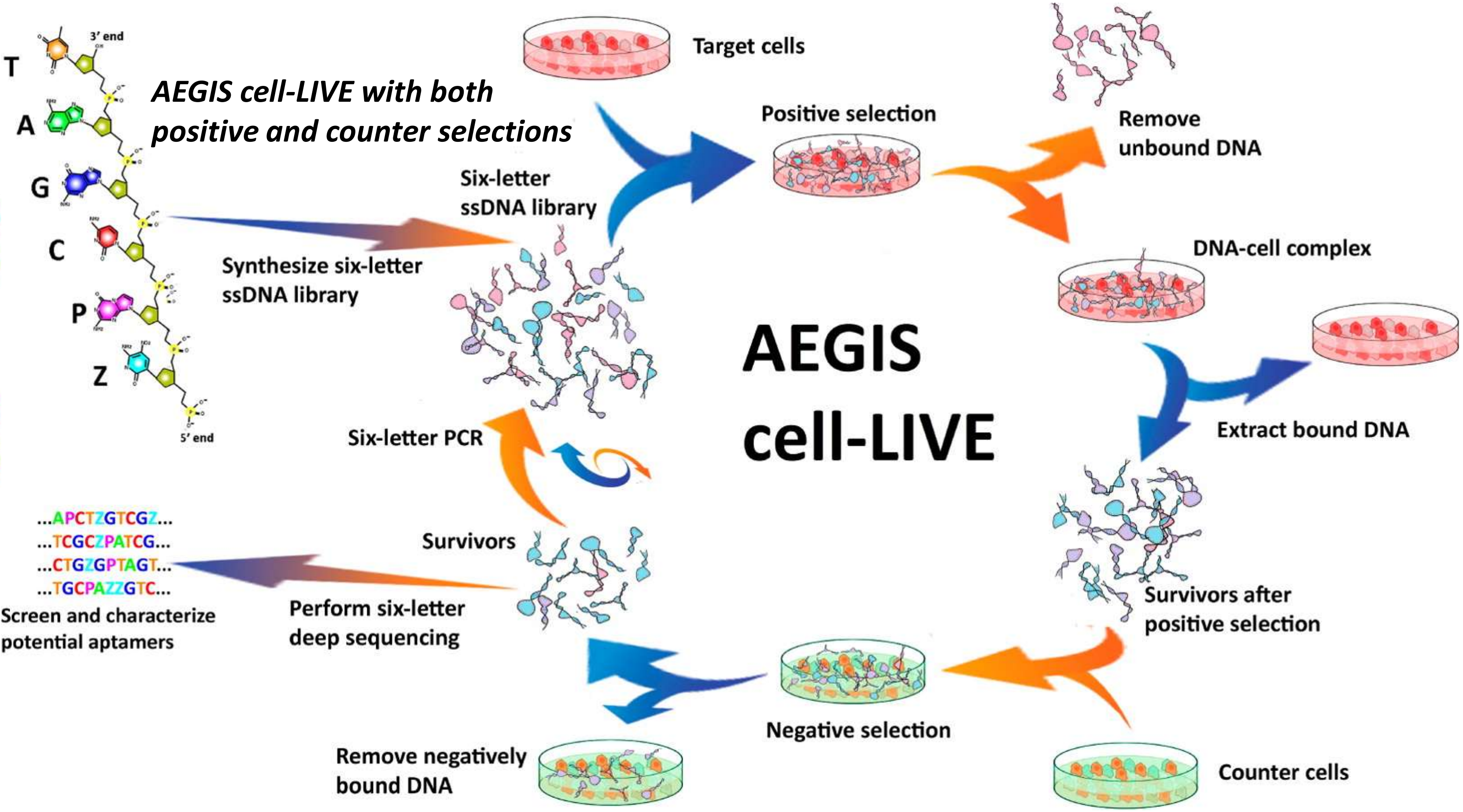
ACGTZP-aptamers

K_d = 24 nM





K_d = 55 nM



AEGIS cell-LIVE with both positive and counter selections

Target cells

Positive selection

Remove unbound DNA

DNA-cell complex

Extract bound DNA

Survivors after positive selection

Counter cells

Negative selection

Remove negatively bound DNA

Survivors

Six-letter PCR

Six-letter ssDNA library

Synthesize six-letter ssDNA library

AEGIS cell-LIVE

Perform six-letter deep sequencing

...APCTZGTCGZ...
...TCGCZPATCG...
...CTGZGPTAGT...
...TGCPAZZGTC...

Screen and characterize potential aptamers

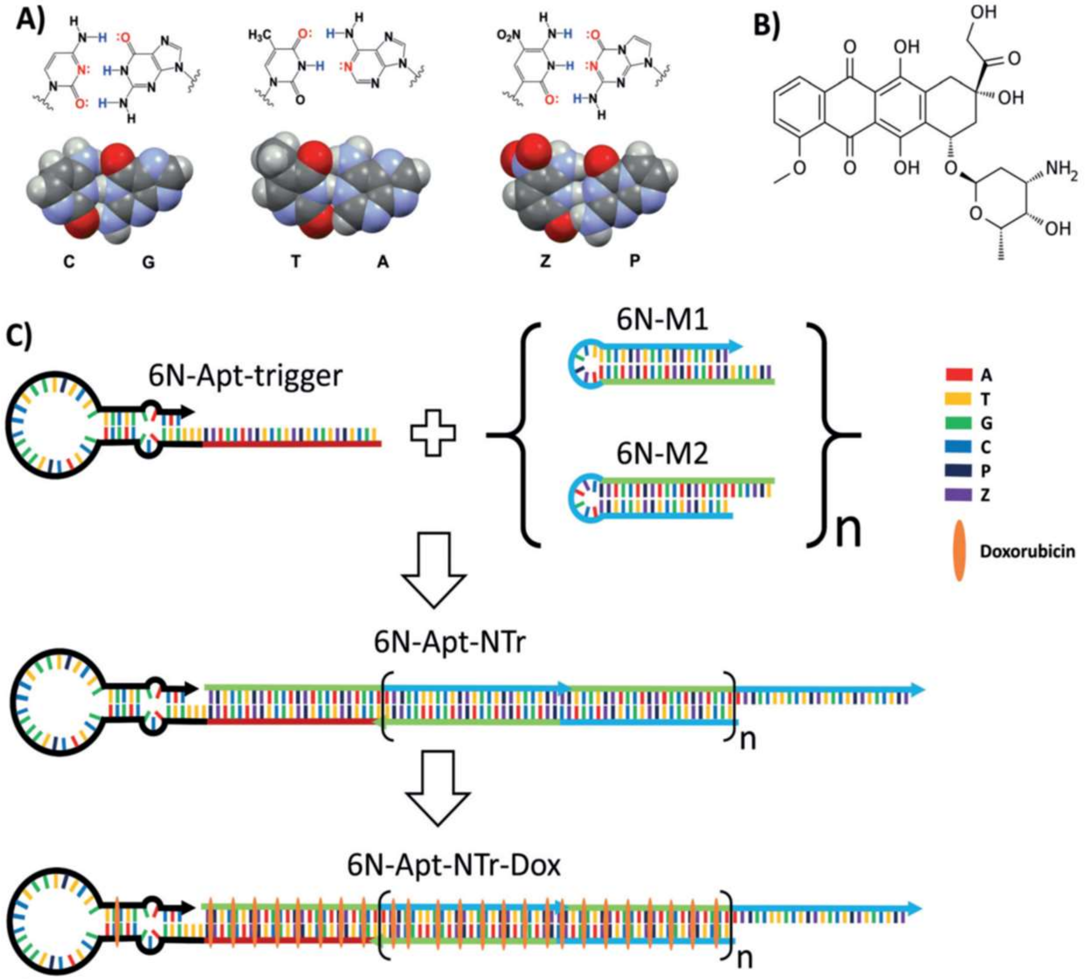
Nucleotides **Z** and **P** were added to a library of oligonucleotides used in a laboratory *in vitro* evolution (LIVE) experiment; the GACTZP library was challenged to deliver molecules that bind selectively to **liver cancer cells**, but not to untransformed **liver cells**. Unlike in classical *in vitro* selection, low levels of mutation allow this system to evolve to create binding molecules not necessarily present in the original library. Over a dozen binding species were recovered. The best had multiple Z and/or P in their sequences.

S. Benner *et al., J. Am. Chem. Soc.,* **2015**, *137,* pp 6734-6737

Aptamer-Nanotrain

The aptamer-nanotrain assembly, charged with doxorubicin, selectively kills liver cancer cells in culture, as the selectivity of the aptamer binding directs doxorubicin into the aptamer-targeted cells. The assembly does not kill untransformed cells that the aptamer does not bind.

The previously selected 6-letter aptamer which binds liver cancer cells



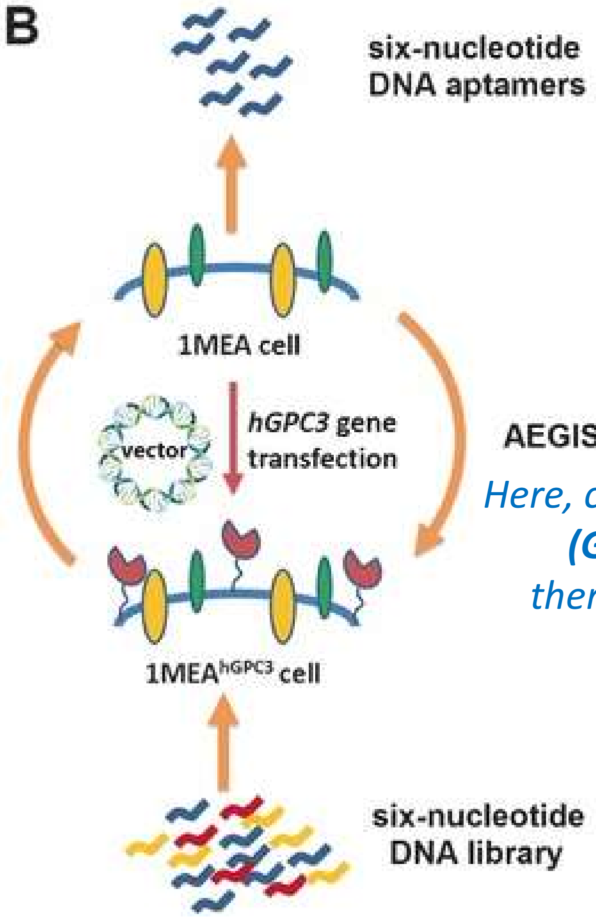
D)

6N-Apt-trigger	5′-TGC TPZ TGC TPZ TGC TPZ ACG AZP TTT GTG ACG CAG CAG CTA CPT GGG CCC TGG TPT CTG TGC TGG ACA C-3′
6N-M1	5′-ZPT CGT PZA GCA PZA GCA PZA GCA AZP GCT TGC TPZ TGC TPZ TGC TPZ-3′
6N-M2	5′-TGC TPZ TGC TPZ TGC TPZ ACG AZP PZA GCA PZA GCA PZA GCA AGC ZPT-3′

This architecture, built with an expanded genetic alphabet, is reminiscent of antibodies conjugated to drugs, which presumably act by this mechanism as well, but with the antibody replaced by an aptamer.

L. Zhang *et al., Angew. Chem., Int.Ed.,* **2020**, *59,* 663-668

ACGTZP-aptamers against a specific protein

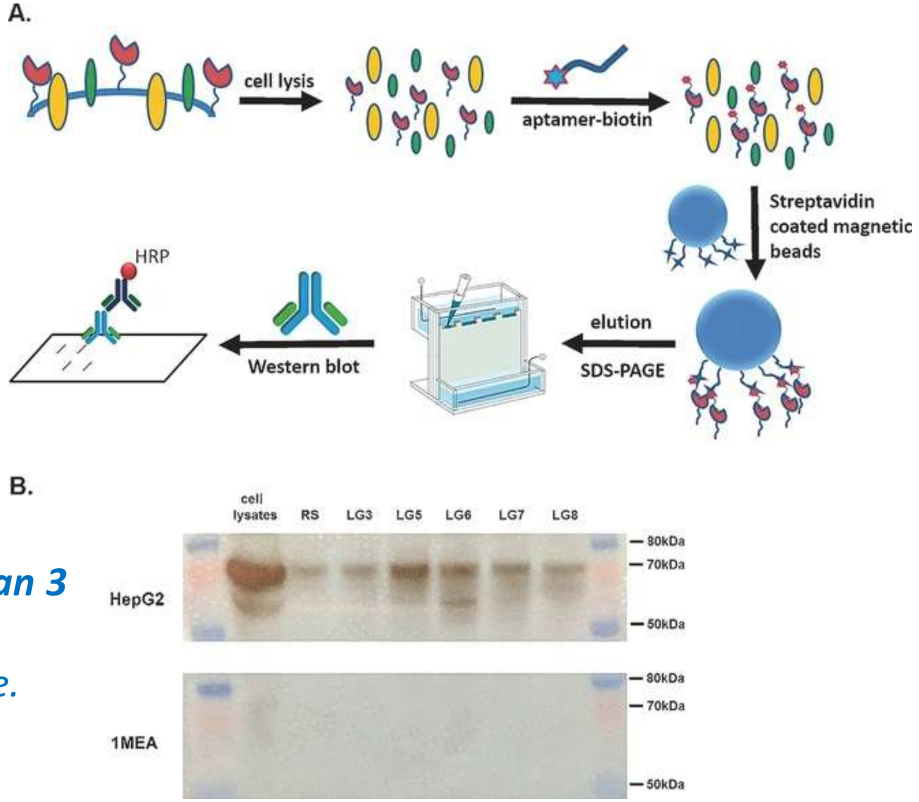


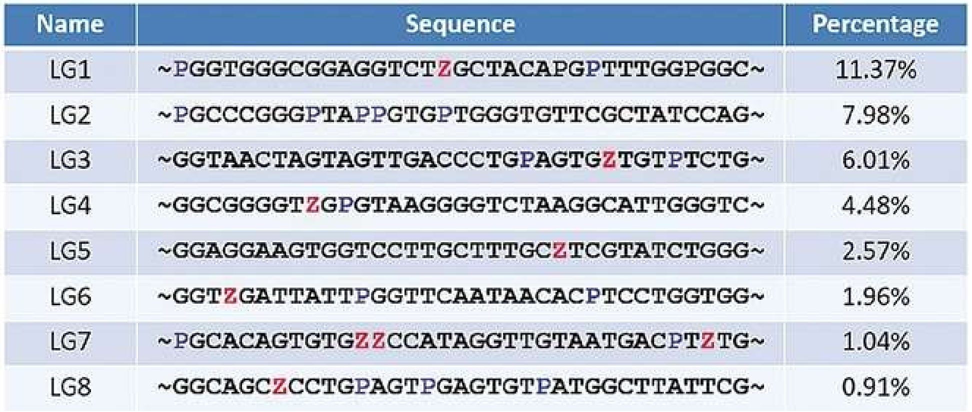
Laboratory in vitro evolution (LIVE) might deliver DNA aptamers that bind proteins expressed on the surface of cells.

*Here, cell engineering was used to place **glypican 3 (GPC3),** a possible marker for liver cancer theranostics, on the surface of a liver cell line.*

Libraries were then built from a six-letter genetic alphabet. With counterselection against non-engineered cells, eight AEGIS-containing aptamers were recovered. Five bound selectively to GPC3-overexpressing cells.

LG5 K_d = 6 nm (without Z – no binding)

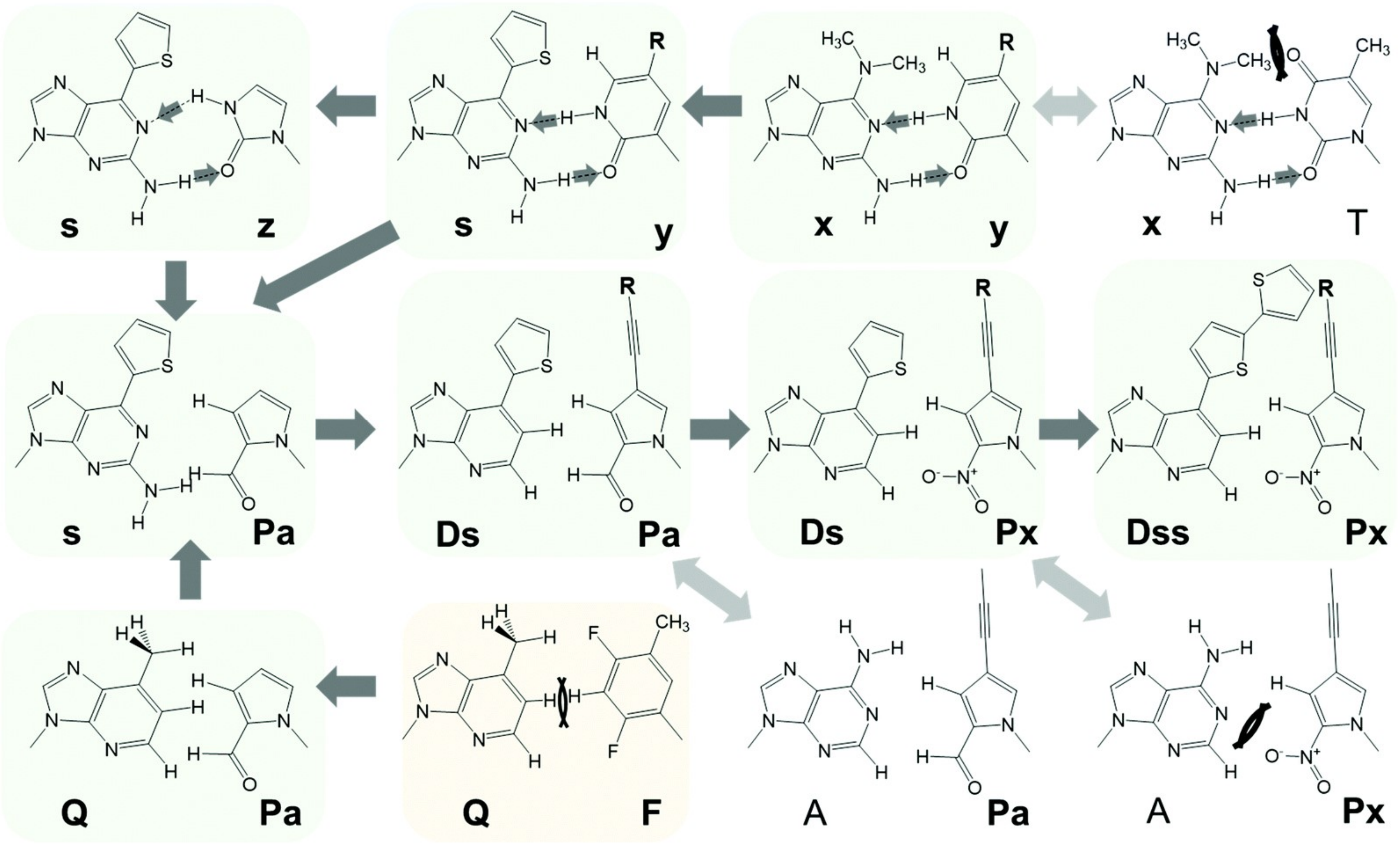




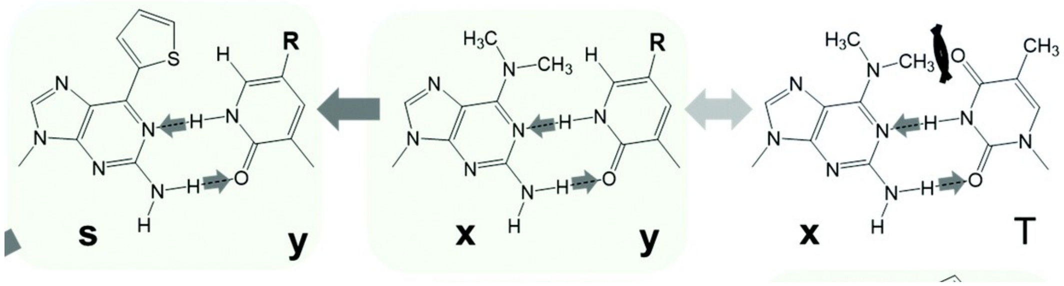
Name	Sequence	Percentage
LG1	~PGGTGGGCGGAGGTCTZGCTACAPGPTTTGGPGGC~	11.37%
LG2	~PGCCCGGGPTAPPGTGPTGGGTGTTCGCTATCCAG~	7.98%
LG3	~GGTAACTAGTAGTTGACCCTGPAGTGZTGTPTCTG~	6.01%
LG4	~GGCGGGGTZGPGTAAGGGGTCTAAGGCATTGGGTC~	4.48%
LG5	~GGAGGAAGTGGTCCTTGCTTTGCZTCGTATCTGGG~	2.57%
LG6	~GGTZGATTATTPGGTTCAATAACACPTCCTGGTGG~	1.96%
LG7	~PGCACAGTGTGZZCCATAGGTTGTAATGACPTZTG~	1.04%
LG8	~GGCAGCZCCTGPAGTPGAGTGTPATGGCTTATTCG~	0.91%

L. Zhang *et al., Angew. Chem., Int.Ed.,* **2016**, *55,* 12372-12375

Steric exclusion and hydrophobic non-natural base pairs



Steric exclusion and hydrophobic non-natural base pairs



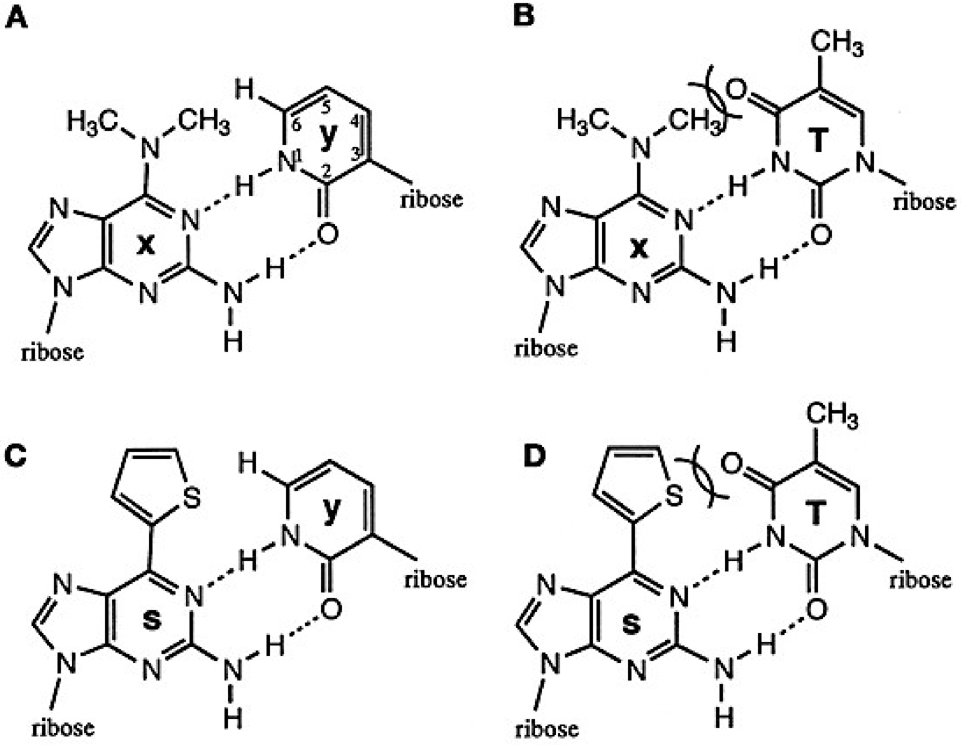
Hirao (2001): the steric hindrance concept to hydrogen-bonded UBPs to exclude the mispairing with natural bases
→ a series of hydrogen-bonded UBPs: **x–y** and **s–y** pairs

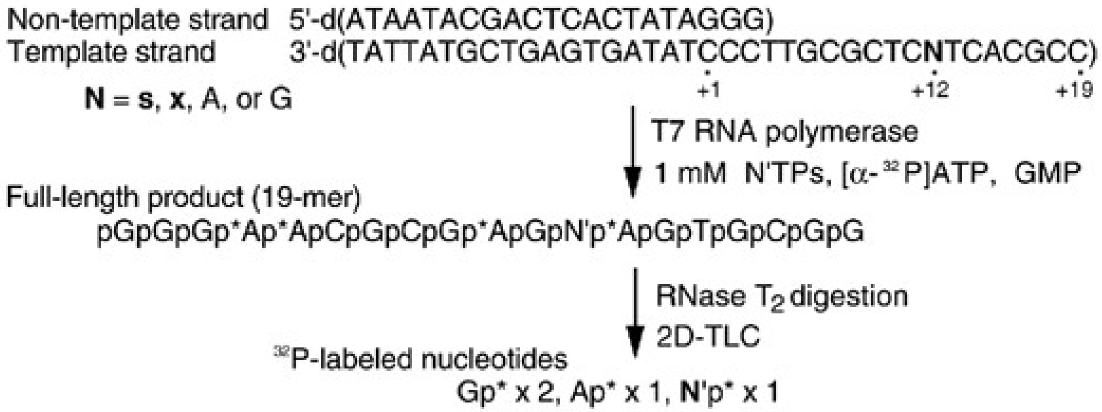
large residues at position 6 of x and s sterically and/or electrostatically clash with the 4-keto group of T, but not with hydrogen of the y base.

x–y and **s–y** pairs function in transcription with T7 RNA polymerase (T7 transcription), and the **y** substrate is incorporated (>96% selectivity **s–y**) site-specifically into RNA transcripts opposite **x** or **s** in DNA templates.

Problem: the **y** base cannot exclude the mispairing with A. Therefore, these UBPs cannot be used in replication

Unnatural aminoacid incorporation using a noncanonical base pair





Non-template strand 5'-d(ATAATACGACTCACTATAGGG)
Template strand 3'-d(TATTATGCTGAGTGATATCCCTTGCGCTC**N**TCACGCC)
+1 +12 +19
N = **s**, **x**, A, or G

T7 RNA polymerase
1 mM N'TPs, [α-^{32}P]ATP, GMP

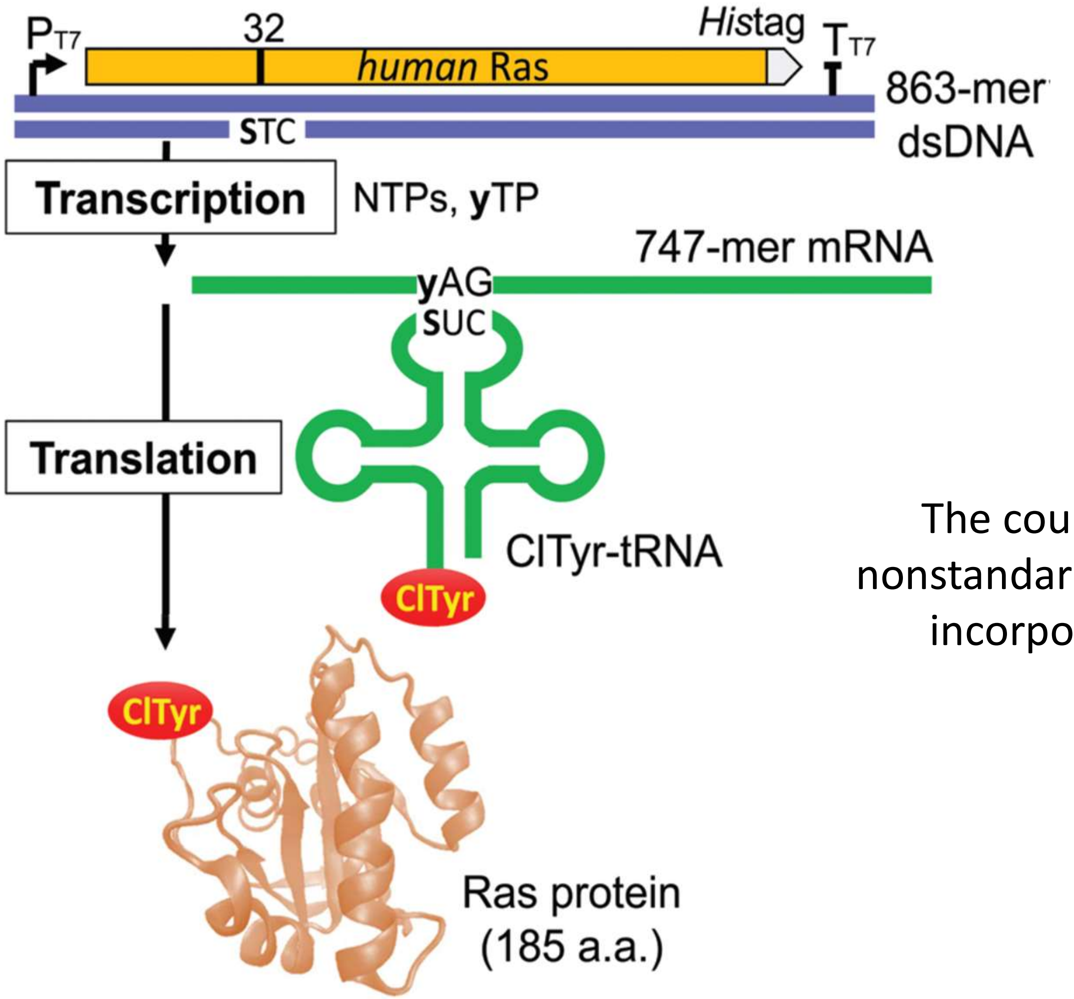
Full-length product (19-mer)
pGpGpGp*Ap*ApCpGpCpGp*ApGpN'p*ApGpTpGpCpGpG

RNase T$_2$ digestion
2D-TLC

^{32}P-labeled nucleotides
Gp* x 2, Ap* x 1, **N**'p* x 1

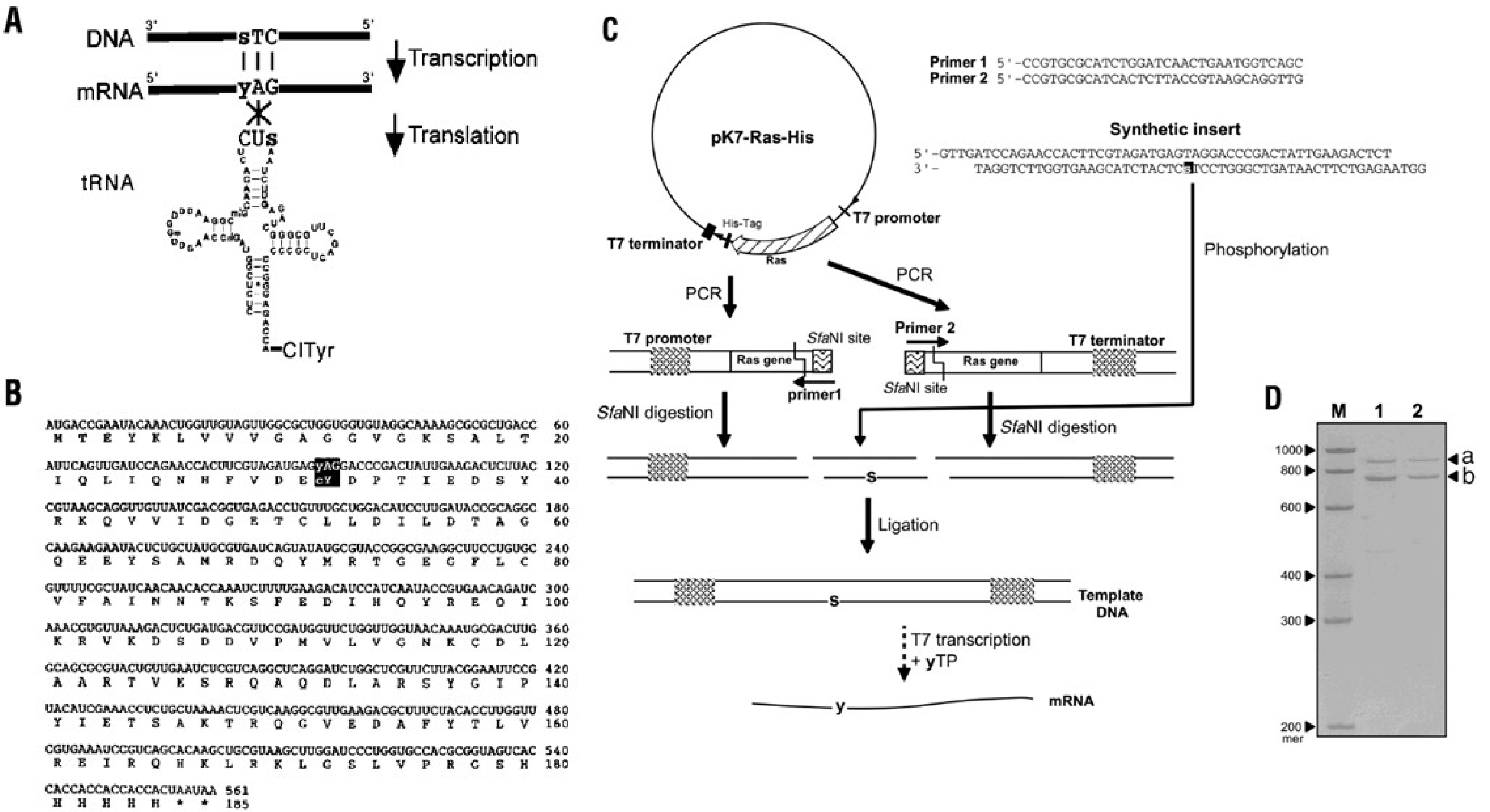
I. Hirao *et al. Nature Biotechnology* **20**, 177–182 (2002)

Unnatural aminoacid incorporation using a noncanonical base pair



The coupled transcription–translation system using the nonstandard codon–anticodon interaction for the site-specific incorporation of 3-chlorotyrosine into the Ras protein.

I. Hirao *et al. Nature Biotechnology* **20**, 177–182 (2002)



```
AUGACCGAAUACAAACUGGUUGUAGUUGGCGCUGGUGGUGUAGGCAAAAGCGCGCUGACC  60
 M  T  E  Y  K  L  V  V  V  G  A  G  G  V  G  K  S  A  L  T   20

AUUCAGUUGAUCCAGAACCACUUCGUAGAUGAGyAGGACCCGACUAUUGAAGACUCUUAC 120
 I  Q  L  I  Q  N  H  F  V  D  E  cY  D  P  T  I  E  D  S  Y   40

CGUAAGCAGGUUGUUAUCGACGGUGAGACCUGUUUGCUGGACAUCCUUGAUACCGCAGGC 180
 R  K  Q  V  V  I  D  G  E  T  C  L  L  D  I  L  D  T  A  G   60

CAAGAAGAAUACUCUGCUAUGCGUGAUCAGUAUAUGCGUACCGGCGAAGGCUUCCUGUGC 240
 Q  E  E  Y  S  A  M  R  D  Q  Y  M  R  T  G  E  C  F  L  C   80

GUUUUCGCUAUCAACAACACCAAAUCUUUUGAAGACAUCCAUCAAUACCGUGAACAGAUC 300
 V  F  A  I  N  N  T  K  S  F  E  D  I  H  Q  Y  R  E  Q  I  100

AAACGUGUUAAAGACUCUGAUGACGUUCCGAUGGUUCUGGUUGGUAACAAAUGCGACUUG 360
 K  R  V  K  D  S  D  D  V  P  M  V  L  V  G  N  K  C  D  L  120

GCAGCGCGUACUGUUGAAUCUCGUCAGGCUCAGGAUCUGGCUCGUUCUUACGGAAUUCCG 420
 A  A  R  T  V  E  S  R  Q  A  Q  D  L  A  R  S  Y  G  I  P  140

UACAUCGAAACCUCUGCUAAAACUCGUCAAGGCGUUGAAGACGCUUUCUACACCUUGGUU 480
 Y  I  E  T  S  A  K  T  R  Q  G  V  E  D  A  F  Y  T  L  V  160

CGUGAAAUCCGUCAGCACAAGCUGCGUAAGCUUGGAUCCCUGGUGCCACGCGGUAGUCAC 540
 R  E  I  R  Q  H  K  L  R  K  L  G  S  L  V  P  R  G  S  H  180

CACCACCACCACCACUAAUAA 561
 H  H  H  H  H  *  *  185
```

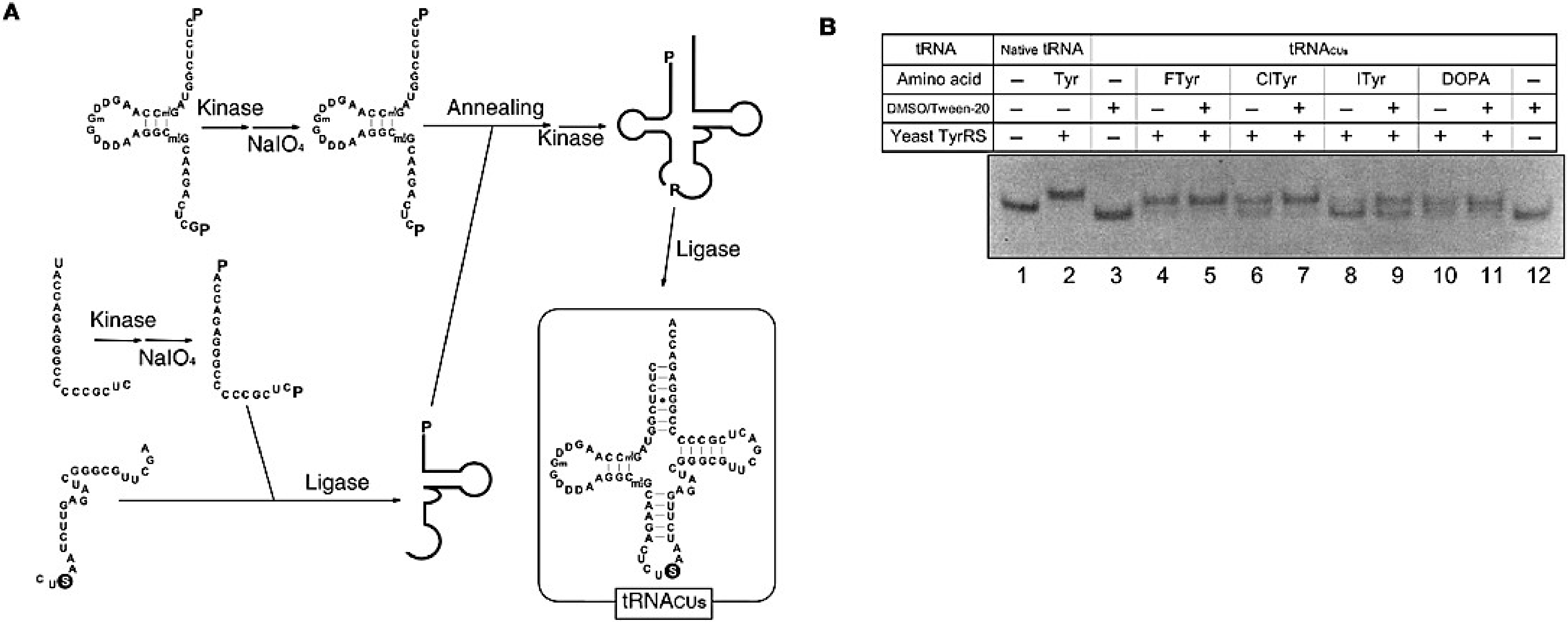
```
Primer 1  5'-CCGTGCGCATCTGGATCAACTGAATGGTCAGC
Primer 2  5'-CCGTGCGCATCACTCTTACCGTAAGCAGGTTG
```

Synthetic insert

```
5'-GTTGATCCAGAACCACTTCGTAGATGAGTAGGACCCGACTATTGAAGACTCT
3'-       TAGGTCTTGGTGAAGCATCTACTCsTCCTGGGCTGATAACTTCTGAGAATGG
```

I. Hirao *et al.* *Nature Biotechnology* **20**, 177–182 (2002)

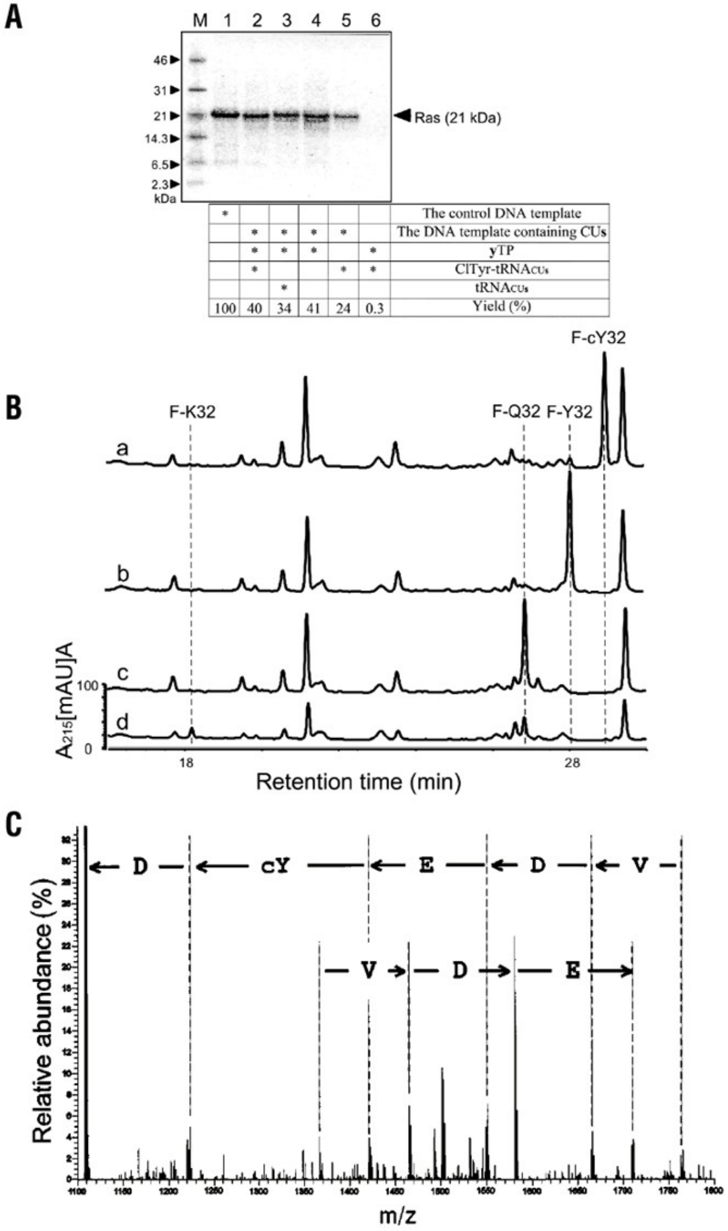
Unnatural aminoacid incorporation using a noncanonical base pair



tRNA	Native tRNA		tRNACUs									
Amino acid	–	Tyr	–	FTyr		ClTyr		ITyr		DOPA		–
DMSO/Tween-20	–	–	+	–	+	–	+	–	+	–	+	+
Yeast TyrRS	–	+	–	+	+	+	+	+	+	+	+	–
	1	2	3	4	5	6	7	8	9	10	11	12

(A) Construction of tRNA$_{CUs}$. (B) Acidic-gel electrophoresis of the products after aminoacylation of *S. cerevisiae* tRNA and tRNA$_{CUs}$ with tyrosine and the 3′-substituted analogs, in the absence or presence of 20% DMSO and 0.25% Tween-20. The upper bands are the aminoacylated tRNAs, and the lower bands are the noncharged tRNAs.

I. Hirao *et al. Nature Biotechnology* **20**, 177–182 (2002)



*						The control DNA template
	*	*	*	*		The DNA template containing CUs
	*	*	*		*	**y**TP
	*			*	*	ClTyr-tRNA$_{CUs}$
		*				tRNA$_{CUs}$
100	40	34	41	24	0.3	Yield (%)

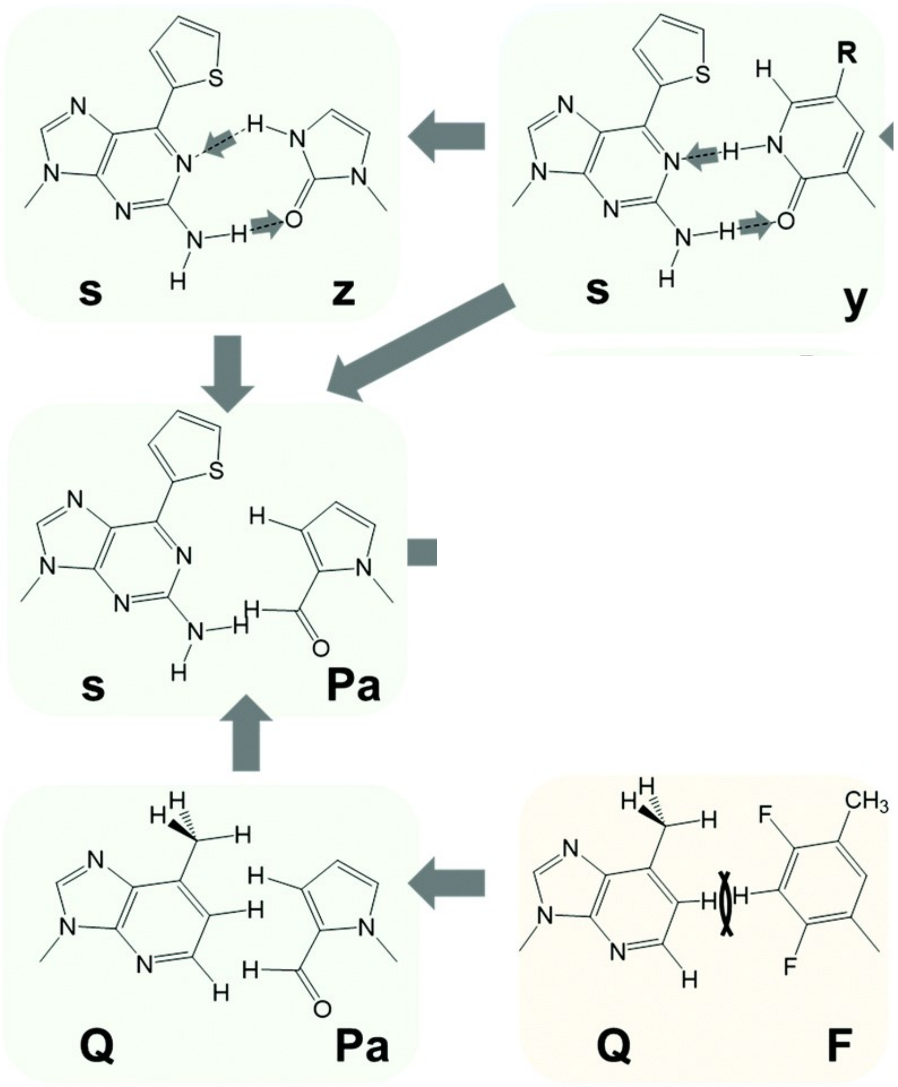
A) Autoradiogram of *in vitro* transcription–translation products labeled with l-[^{14}C]leucine. The conditions are noted at the bottom of each lane.

B) LC patterns of the products digested by Lys-C. Chart a shows the products obtained from the reaction in the presence of the DNA template containing the CT**s** sequence, **y**TP, and ClTyr-tRNA$_{CU\mathbf{s}}$; chart b shows those obtained in the presence of the standard DNA template and Tyr; chart c shows those obtained in the presence of the DNA template containing CT**s** but in the absence of **y**TP and tRNA$_{CU\mathbf{s}}$; and chart d shows those obtained in the presence of the DNA template containing CT**s** and **y**TP but in the absence of tRNA$_{CU\mathbf{s}}$.

C) Tandem mass spectrum of the F-cY32 fragment. The partial sequence, VDE**cY**D, of F-cY32 was confirmed from the ion series

I. Hirao *et al. Nature Biotechnology* **20**, 177–182 (2002)

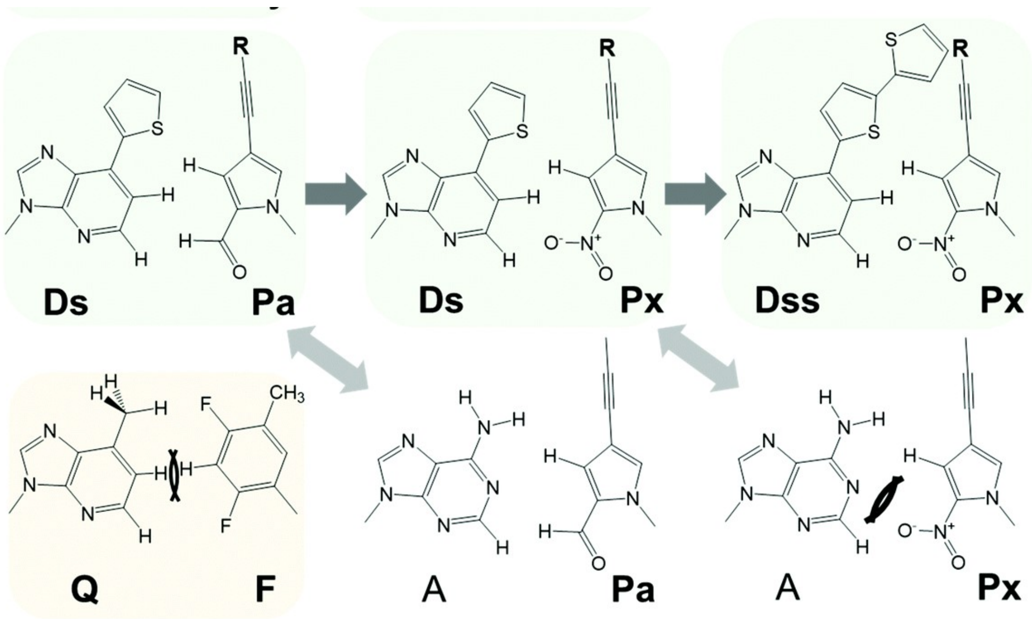
Steric exclusion and hydrophobic non-natural base pairs



Kool (1998): non-hydrogen-bonded **Z–F** pair (an isostere of A–T) (Z is 4-methylindole, F base is 2,4-difluorotoluene.), but lack of Z and F interactions with polymerases.
Kool (1999): Q-F pair - shape-complementarity rather than the hydrogen-bond interactions, ➔a new strategy of UBP development using hydrophobic UBs without any hydrogen-bond interactions between the pairing bases.

Hirao (2003): the Q–Pa pair to avoid steric clashes.
Hirao (2007): Pa could be used as a pairing partner of the s base, and the s substrate was incorporated specifically into RNA opposite Pa in templates by T7 transcription.
Hirao (2004): The s–z pair increased the s incorporation selectivity opposite z in T7 transcription, as compared to that opposite y in the s–y pair

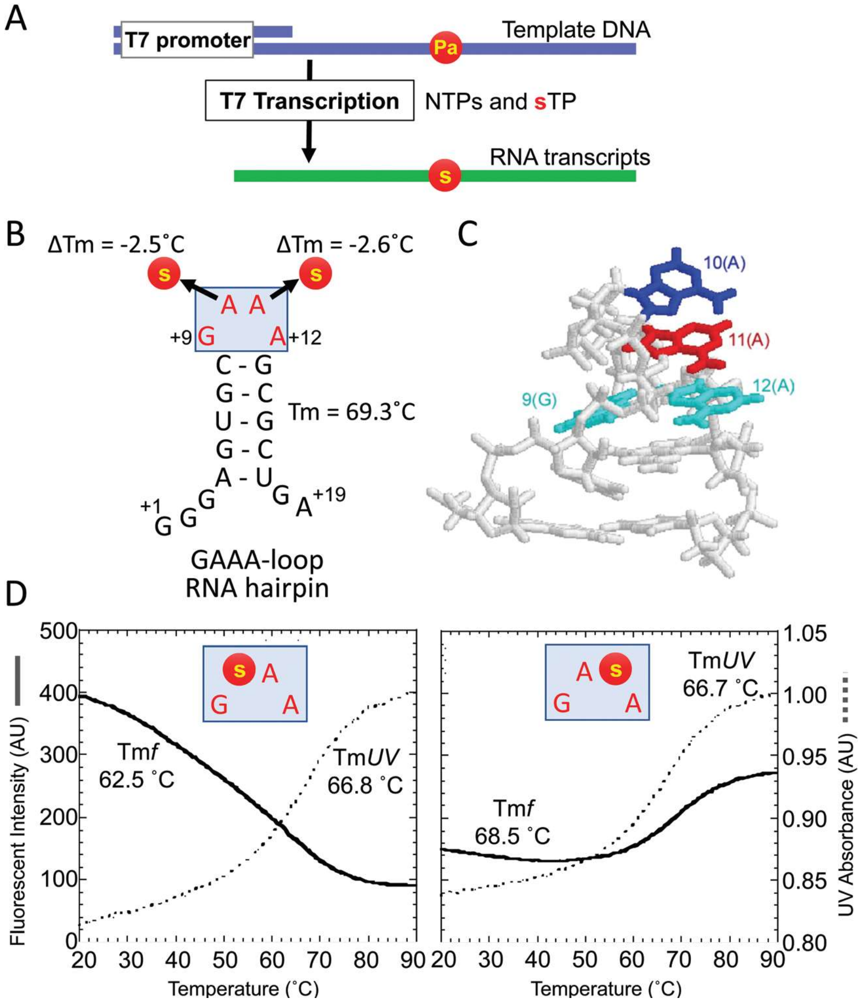
Steric exclusion and hydrophobic non-natural base pairs

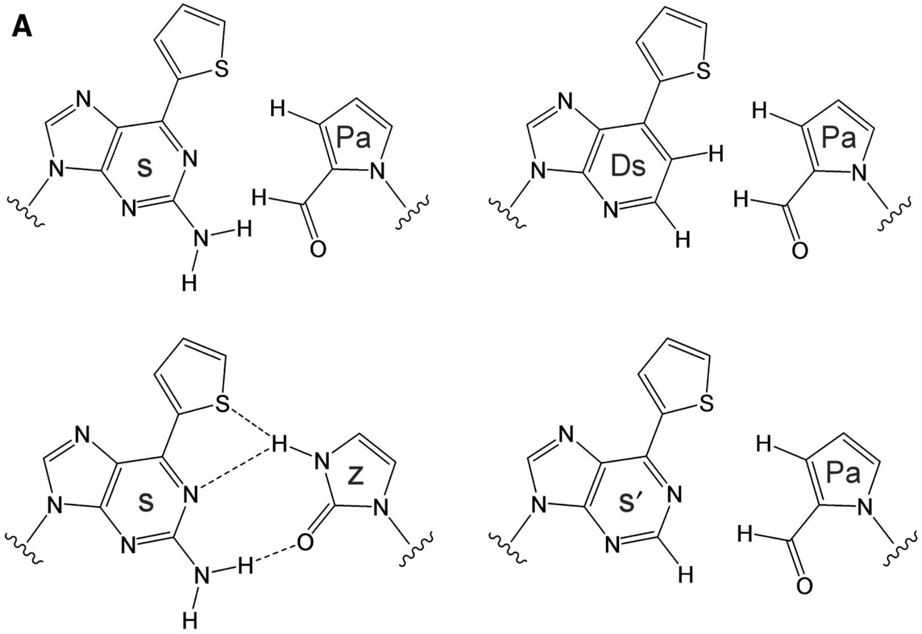


Hirao (2006) the hydrophobic **Ds** base (improved **s**). The **Ds–Pa** pair exhibited high incorporation efficiency in replication. **Problems**: the non-cognate Ds–Ds and A–Pa pairings (solved by using γ-amido-triphosphates of Ds and A, which significantly reduced their mispairing with Ds and Pa - ~99.9% selectivity per cycle in PCR).

Hirao (2007): the **Ds–Px** pair - electrostatic clash with the 1-nitrogen of A (no A–Px pairing), The additional propynyl group reduces the Ds–Ds self-pairing.
The Ds–Px pair exhibits high fidelity and high efficiency in PCR (>99.9% per cycle).

Fluorescent probing for RNA molecules by an unnatural base-pair system



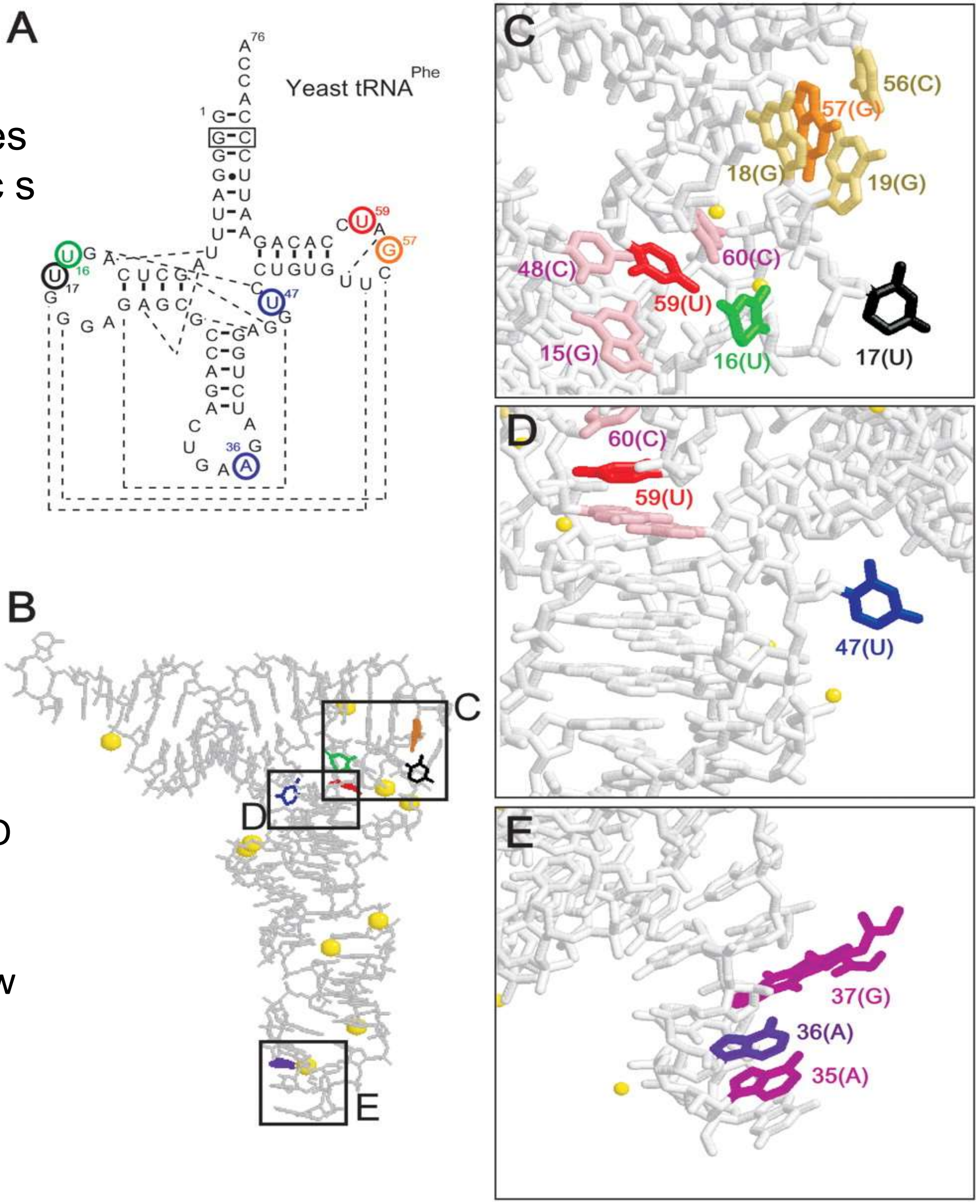


(B) Incorporation of the fluorescent **s** base into GNRA hairpins. (*left*) The secondary structure of the RNA hairpin with a GAAA loop. The second A and third A in the loop are shown in blue and red, respectively. *(right)* The 3D structure of the GAAA-loop hairpin

M. Kimoto, T. Mitsui, Y. Harada, A. Sato, S. Yokoyama and I. Hirao, *Nucleic Acids Res.*, **2007**, *35*, 5360–5369

The „s” incorporation sites in yeast $tRNA^{Phe}$

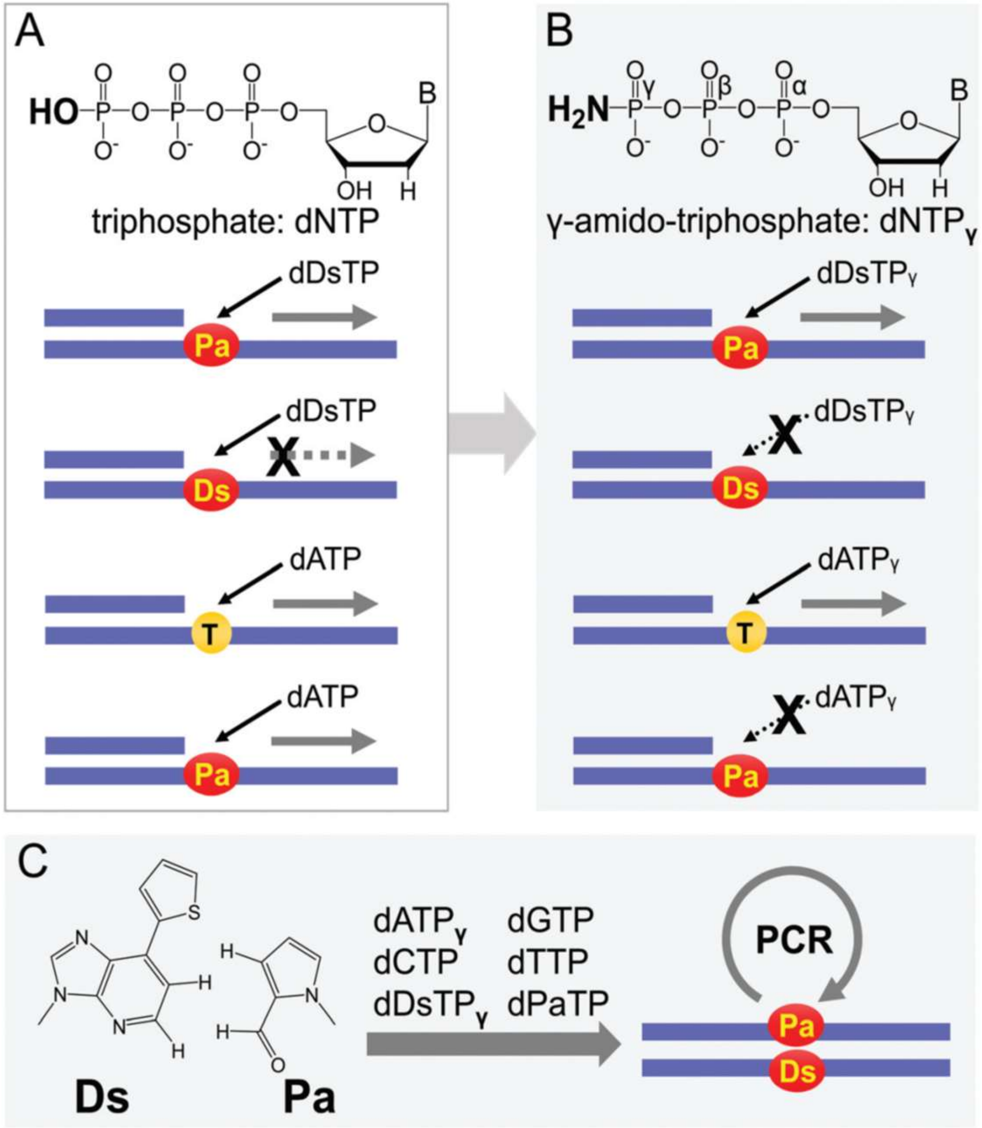
The fluorescent intensity of „s” in RNA molecules changes according to the structural environment. The site-specific s labeling of RNA hairpins and tRNA molecules provided characteristic fluorescent profiles, depending on the labeling sites, temperature and Mg^{2+} concentration.



(**A**) The secondary structure of the original tRNA transcript. The positions substituted with **s** are circled. The broken lines show base–base interactions for the 3D structure.
(**B**–**E**) The deep-colored bases were substituted with **s**, which stacks with the light-colored bases, and the yellow spheres represent Mg^{2+}.

M. Kimoto, T. Mitsui, Y. Harada, A. Sato, S. Yokoyama and I. Hirao, *Nucleic Acids Res.*, **2007**, *35*, 5360–5369

Steric exclusion and hydrophobic non-natural base pairs



The Ds–Pa pair for faithful replication in combination with γ-amidotriphosphates.

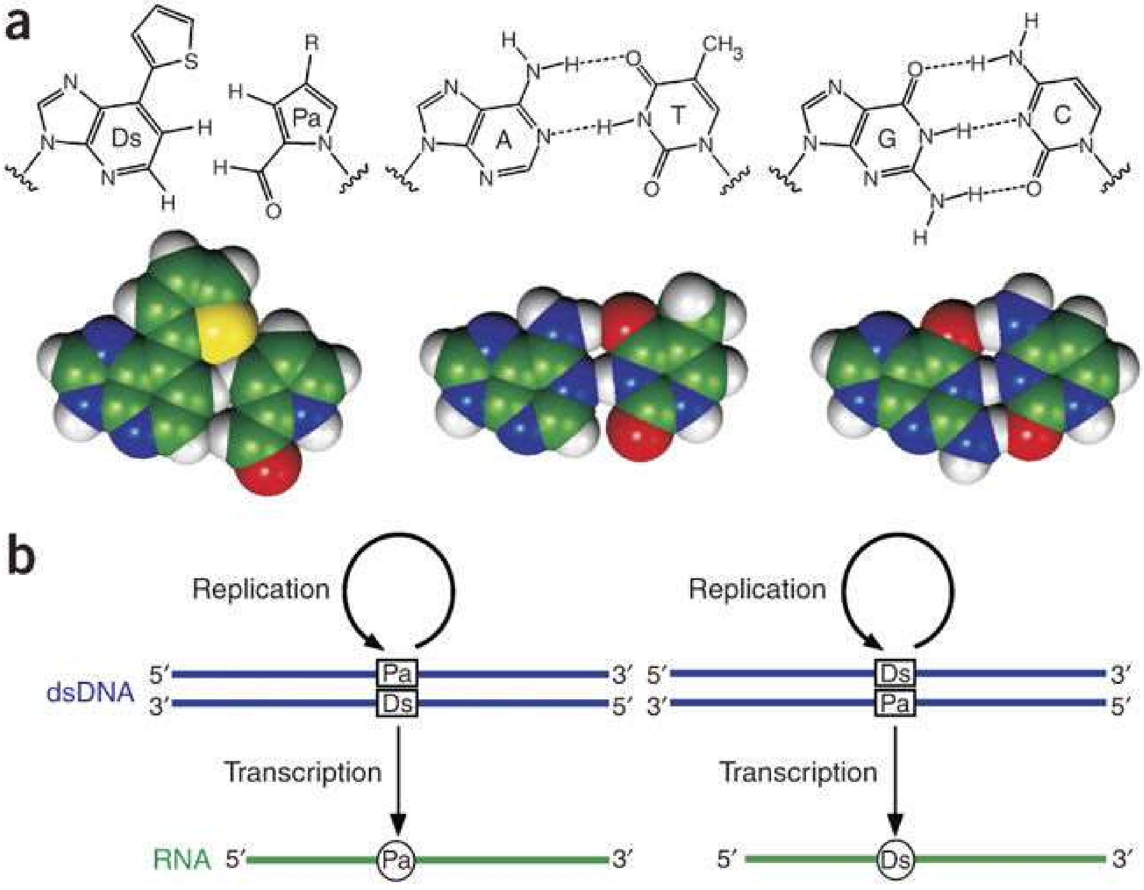
(A) The substrates, dDsTP and dATP, are incorporated opposite Pa and T in the templates, respectively. However, dDsTP and dATP are also misincorporated opposite Ds and Pa, respectively. The Ds incorporation opposite Ds inhibits further primer extension.

(B) The γ-amido-triphosphates, dDsTPγ and dATPγ, reduce the Ds–Ds and A–Pa mispairings in replication.

(C) The combination of the usual triphosphates (dG/C/T/PaTP) and the γ-amido-triphosphates (dDs/ATPγ) enables the faithful PCR amplification of 6-letter DNA

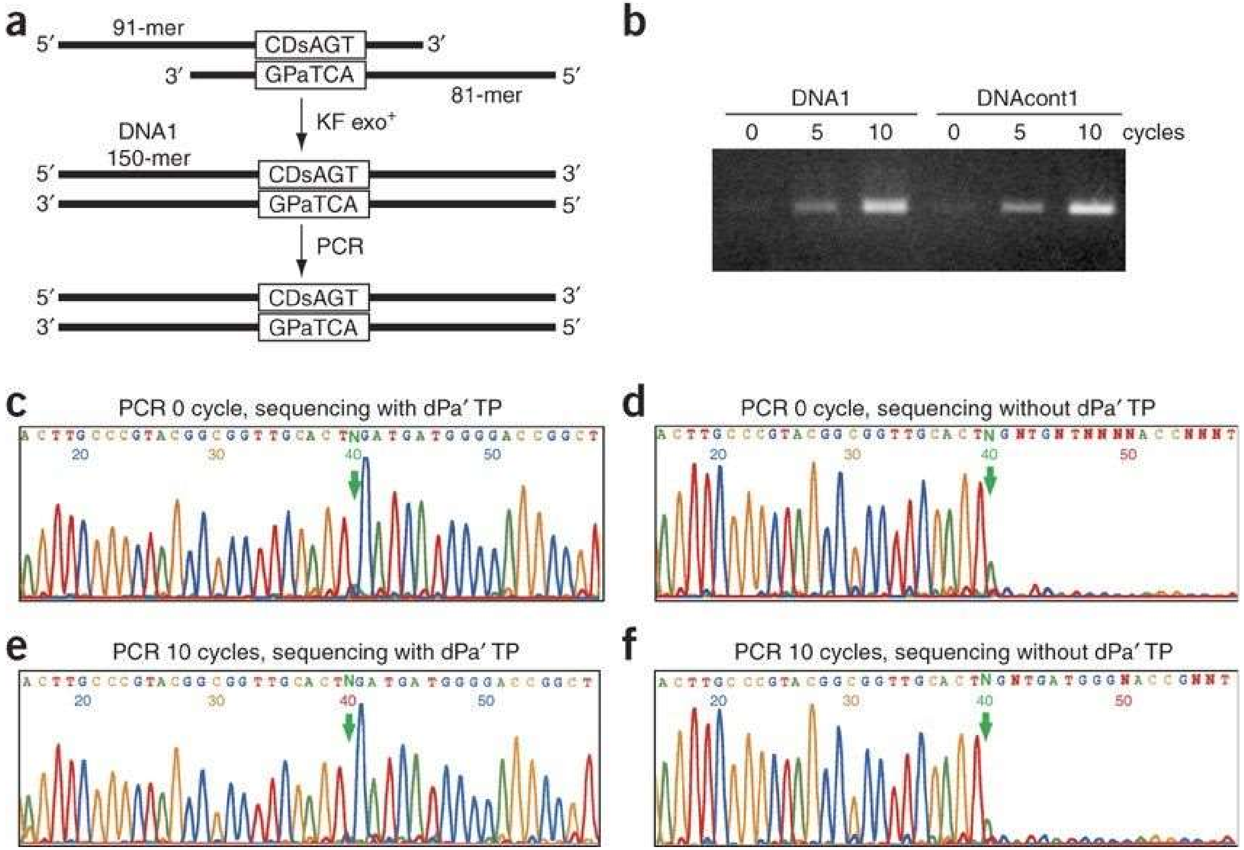
M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

Replication and transcription of the hydrophobic UBP



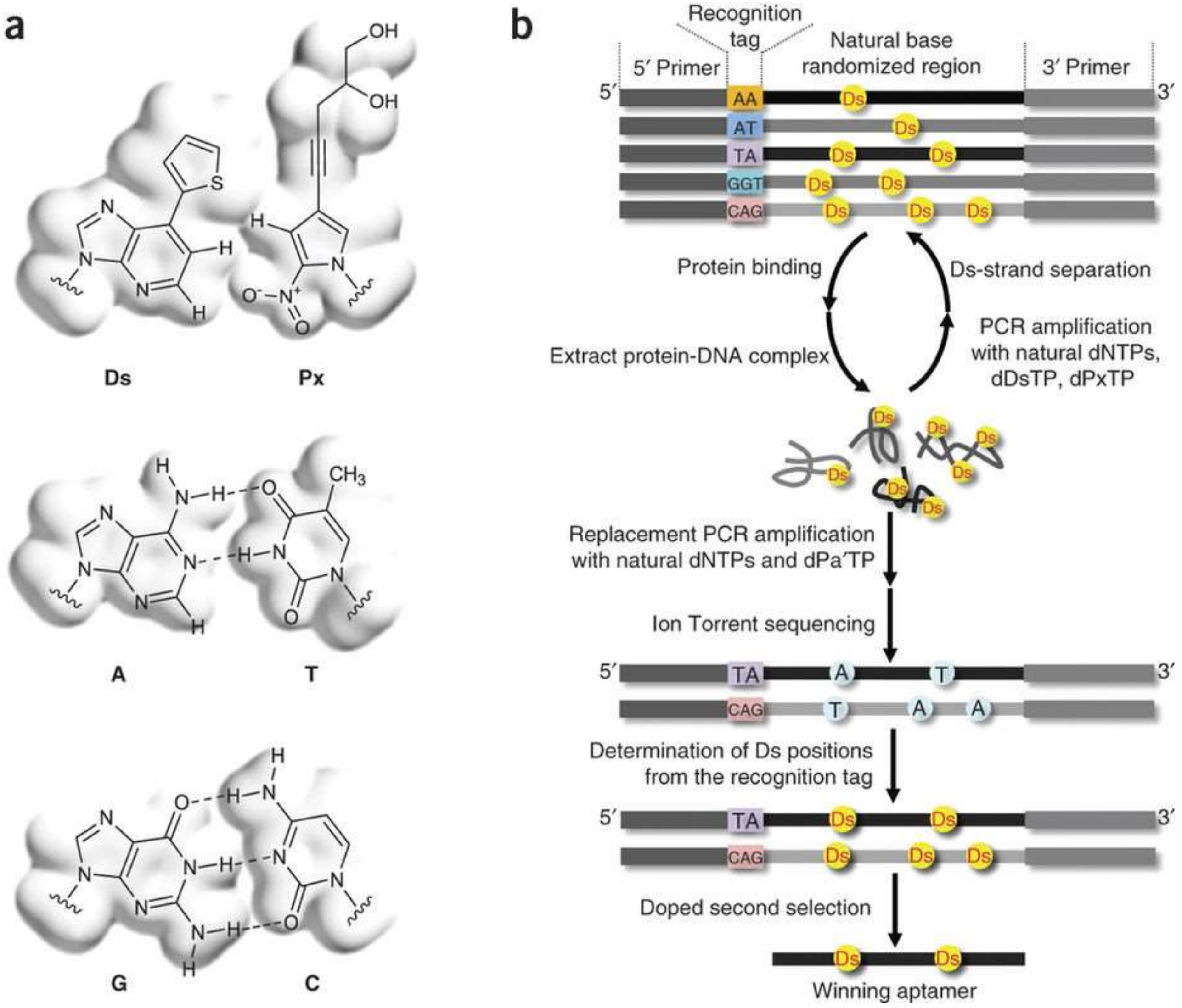
(**a**) The unnatural Ds-Pa and natural base pairs (Pa: R = H, Pa': R = C≡C–CH_3). (**b**) The unnatural base pair system that functions in PCR amplification, primer extension, DNA sequencing and T7 transcription. Original DNA templates were prepared by chemical synthesis and ligation, and were amplified by PCR with unmodified dNTPs (dPaTP, dGTP, dCTP and dTTP) and modified $dNTP_N$s (γ-amidotriphosphates, $dDsTP_N$ and $dATP_N$). RNA molecules containing Pa or modified Pa bases at specific positions were transcribed from DNA templates containing Ds in the template strands, by T7 RNA polymerase with PaTP (or modified PaTP) and the natural NTPs (left). RNA molecules containing Ds were transcribed from DNA templates containing Pa in the template strands, with DsTP and the natural NTPs (right).

I. Hirao *et al. Nature Methods* **2006**, *3,* 729-735



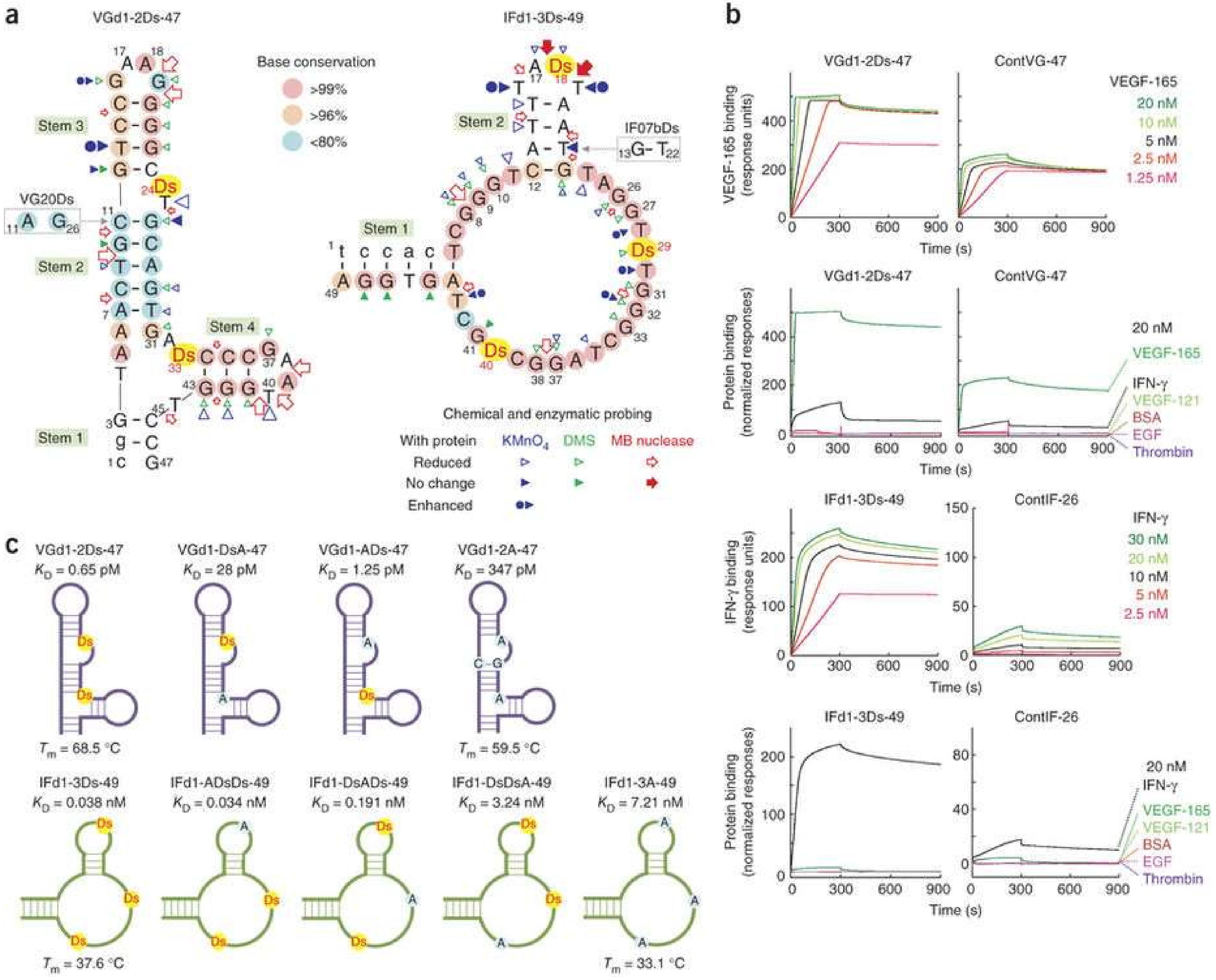
(**a**) The double stranded DNA fragment (150-mer, DNA1) was prepared by primer extension using chemically synthesized DNA fragments (91-mer and 81-mer) containing Ds and Pa. (**b**) Agarose-gel analysis of original DNA fragments (0 cycle) and PCR products after 5 and 10 cycles of amplification. For DNA1, PCR was performed with 0.04 unit/µl Vent DNA polymerase and the following cycle conditions: 0.5 min at 94 °C, 0.5 min at 45 °C and 4 min at 65 °C. For DNAcont1 consisting only of the natural bases, PCR was performed with 0.01 unit/µl Vent DNA polymerase with the following cycle conditions: 0.5 min at 94 °C, 0.5 min at 45 °C and 1 min at 72 °C. (**c**–**f**) DNA sequencing, in the presence (**c,e**) or absence (**d**,**f**) of dPa′TP, of the original DNA1 (**c**,**d**) and PCR-amplified DNA1 after 10 cycles using the unnatural base pair system (**e**,**f**).

Ds-Px noncanonical base pair – High-affinity DNA aptamers



I. Hirao *et al. Nature Biotechnology* **2013**, *31,* 453-458

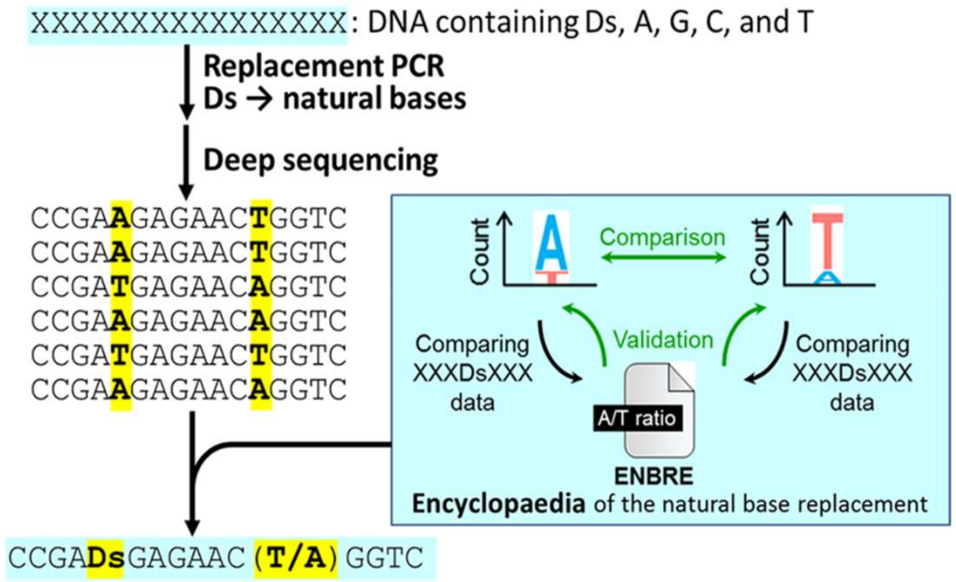
Targets:
VEGF-165
and IFN-γ



Characterizations and binding affinities of anti–VEGF-165 aptamer (VGd1-2Ds-47) and anti–IFN-γ aptamer (IFd1-3Ds-49)

I. Hirao *et al. Nature Biotechnology* **2013**, *31,* 453-458

Ds-Pa – need for deep sequencing in the amplification process



XXXXXXXXXXXXXXXX: DNA containing Ds, A, G, C, and T

Replacement PCR
Ds → natural bases

Deep sequencing

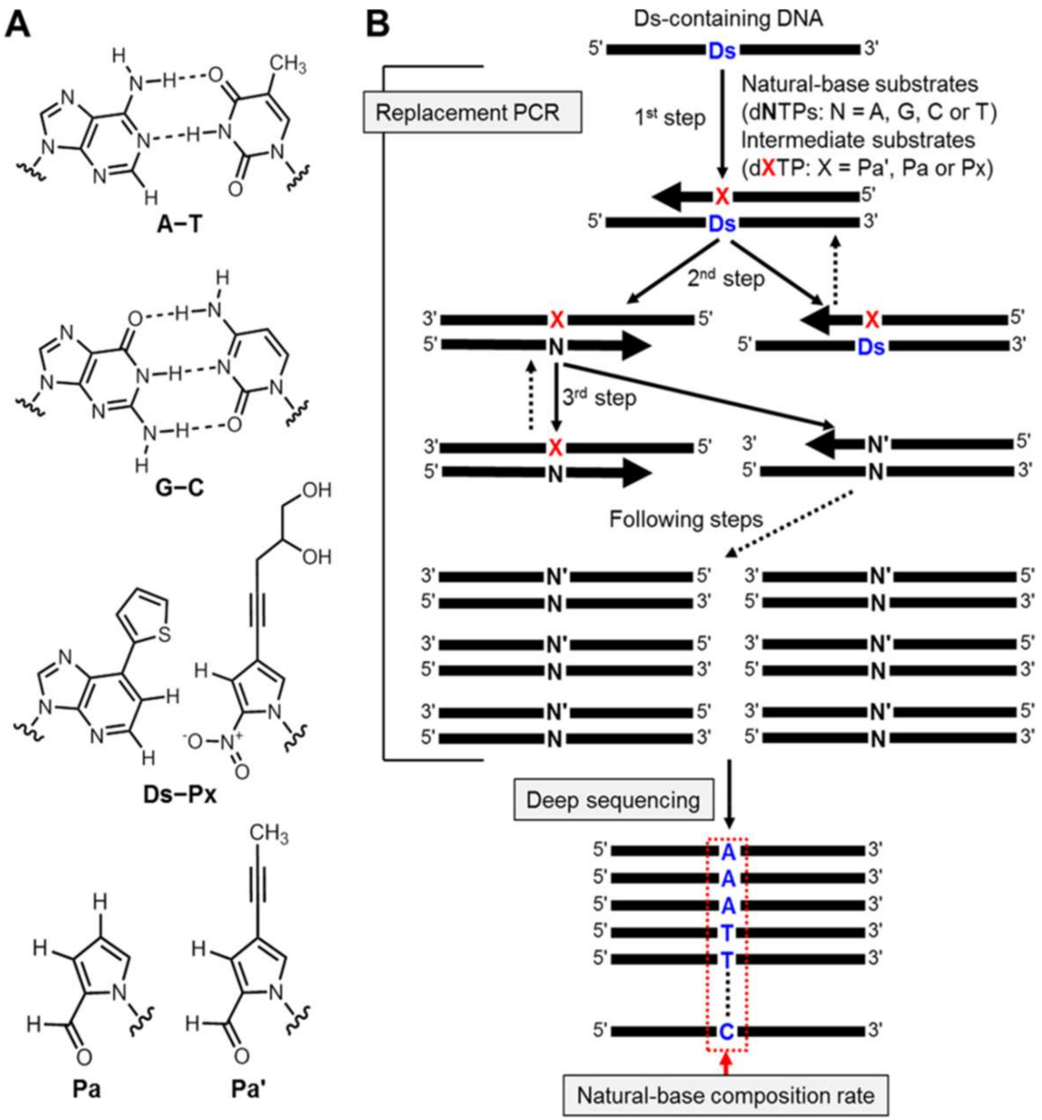
CCGA**A**GAGAAC**T**GGTC
CCGA**A**GAGAAC**T**GGTC
CCGA**T**GAGAAC**A**GGTC
CCGA**A**GAGAAC**A**GGTC
CCGA**T**GAGAAC**T**GGTC
CCGA**A**GAGAAC**A**GGTC

Comparing XXXDsXXX data

Comparing XXXDsXXX data

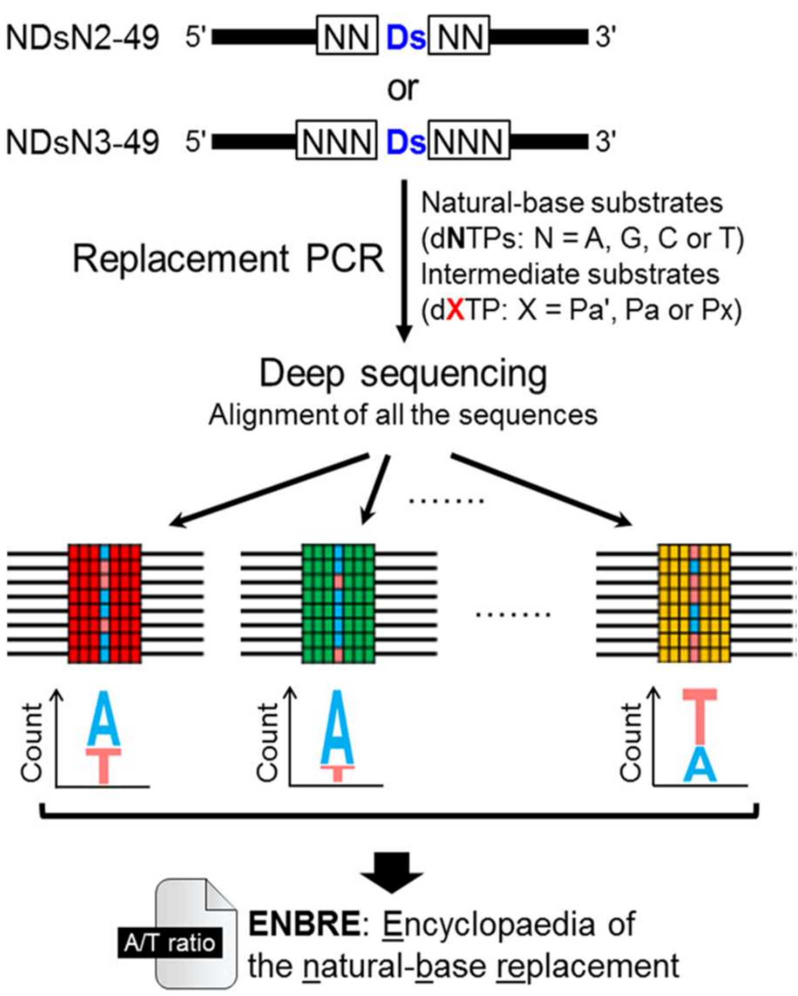
ENBRE
Encyclopaedia of the natural base replacement

CCGA**Ds**GAGAAC(**T/A**)GGTC



K. Hamashima, I. Hirao *et al. ACS Synth. Biol.* **2019**, *8,* 1401-1410

Ds-Pa – deep sequencing



NDsN2-49 5' NN **Ds** NN 3'

or

NDsN3-49 5' NNN **Ds** NNN 3'

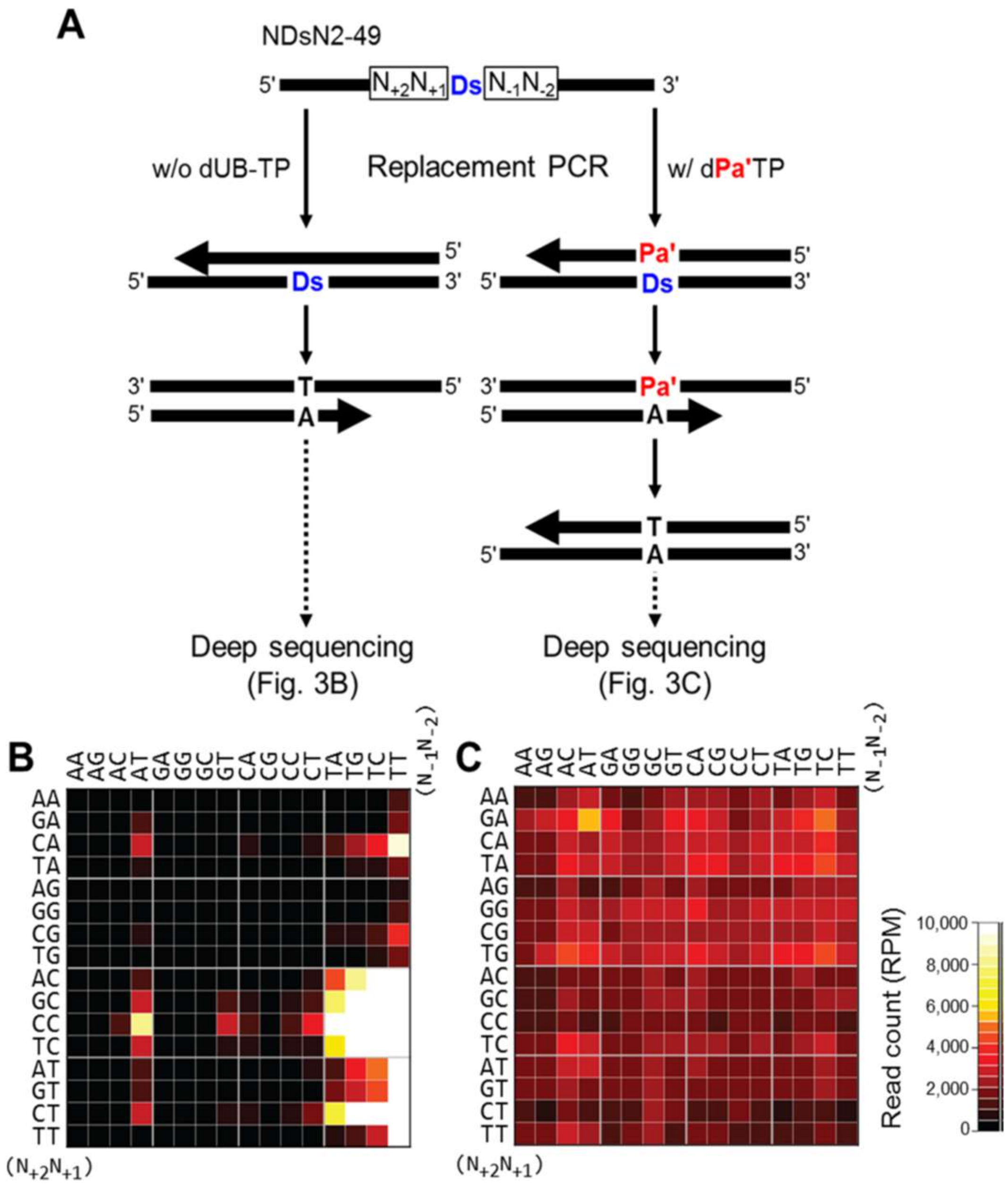
Replacement PCR

Natural-base substrates
(d**N**TPs: N = A, G, C or T)
Intermediate substrates
(d**X**TP: X = Pa', Pa or Px)

Deep sequencing
Alignment of all the sequences

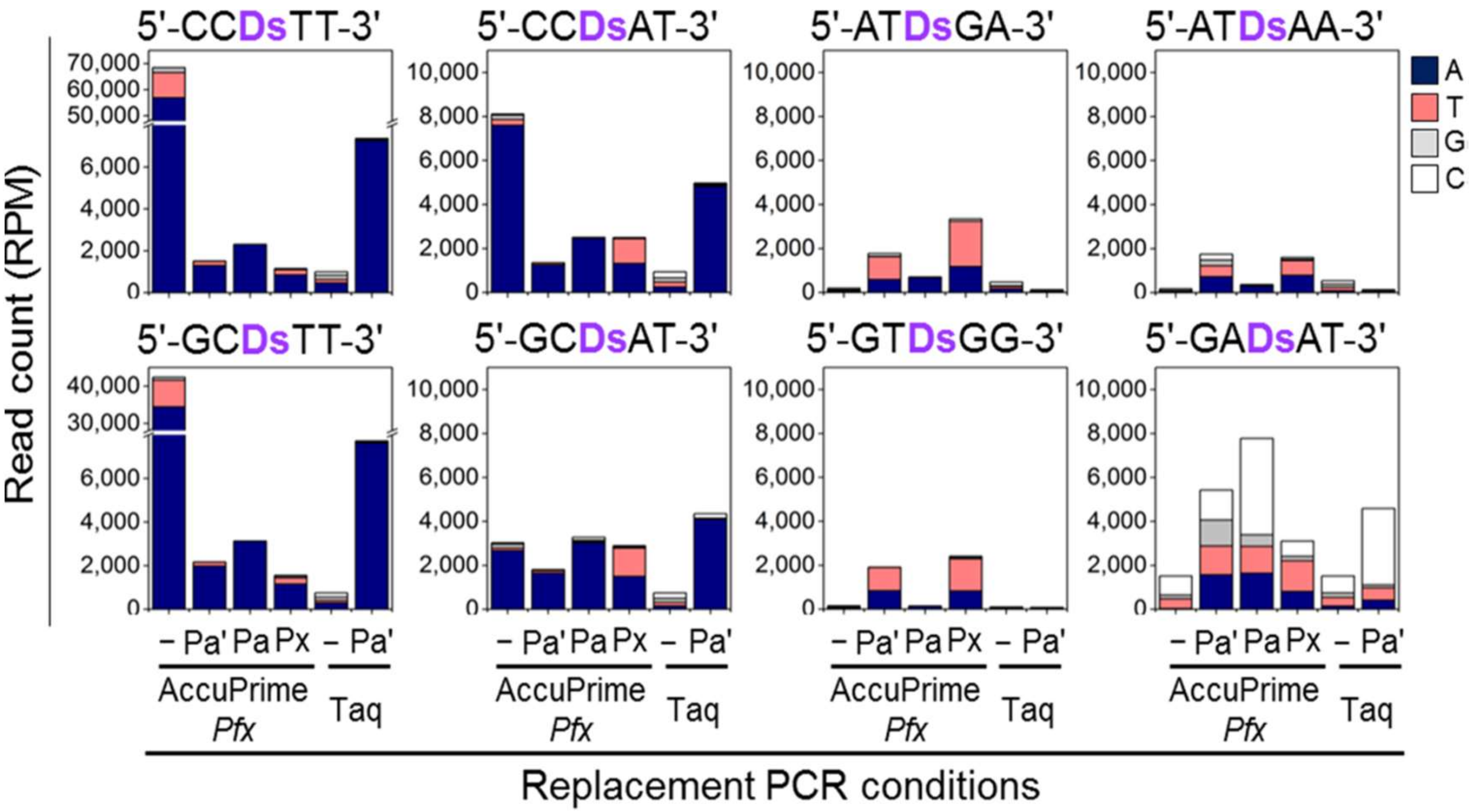
A/T ratio

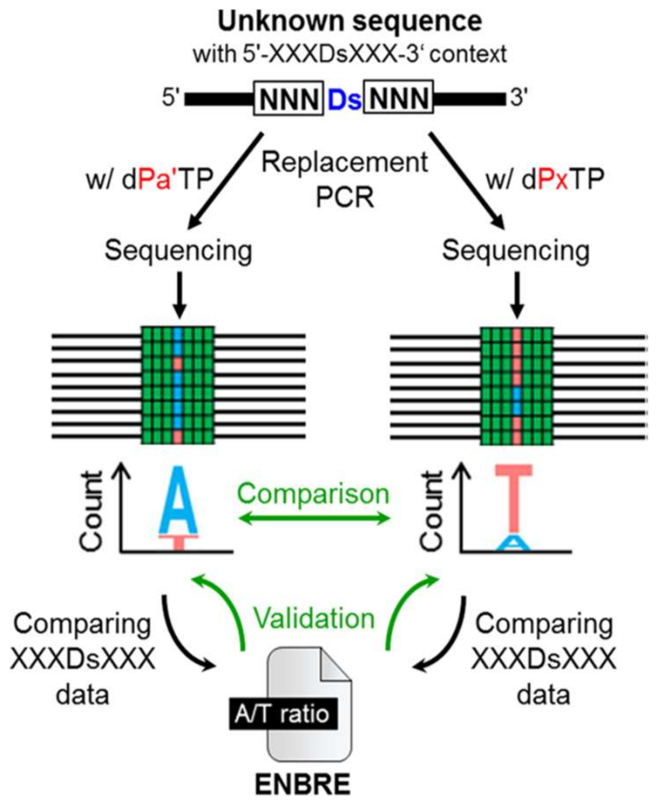
ENBRE: Encyclopaedia of the natural-base replacement



K. Hamashima, I. Hirao *et al. ACS Synth. Biol.* **2019**, *8,* 1401-1410

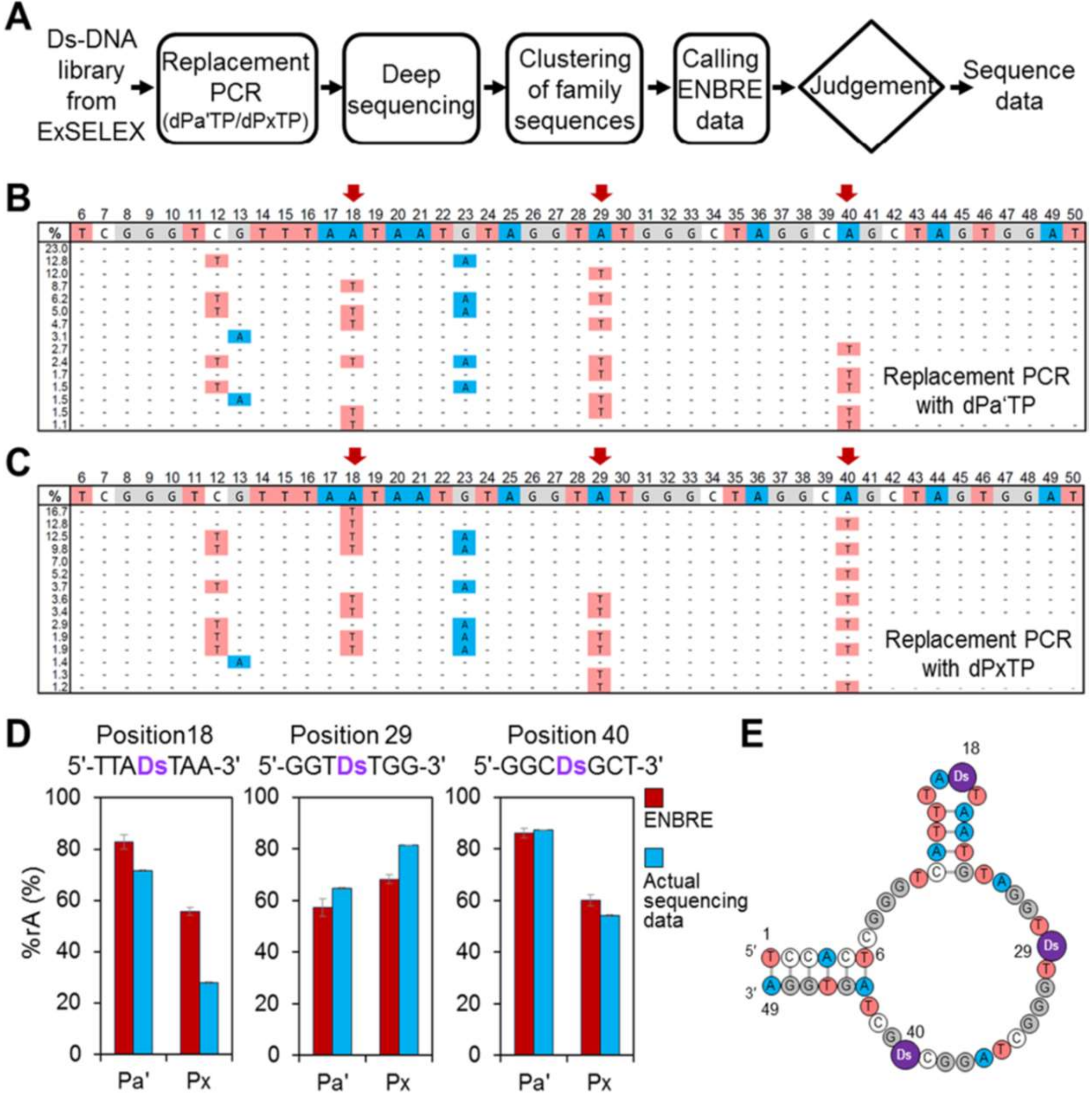
Ds-Pa – deep sequencing





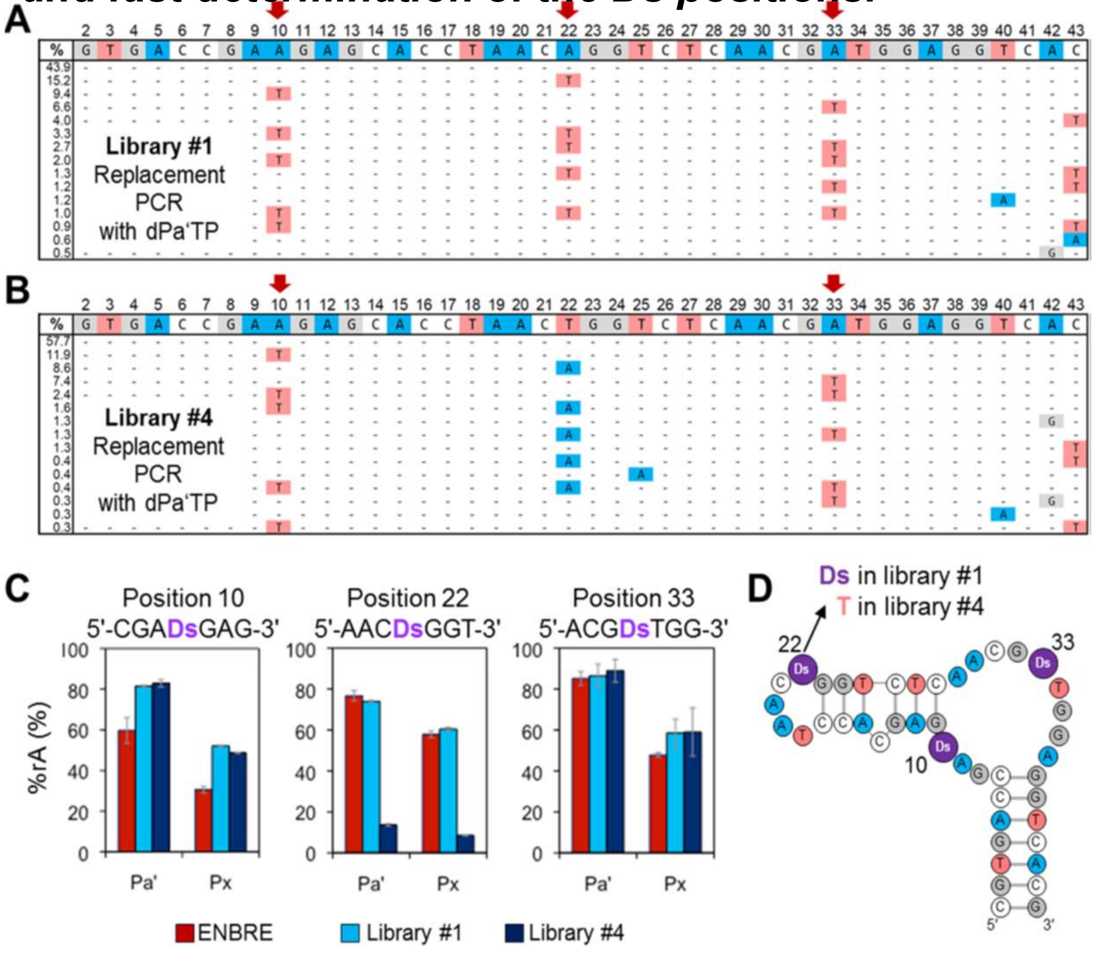
K. Hamashima, I. Hirao *et al. ACS Synth. Biol.* **2019**, *8,* 1401-1410

Ds-Pa – deep sequencing



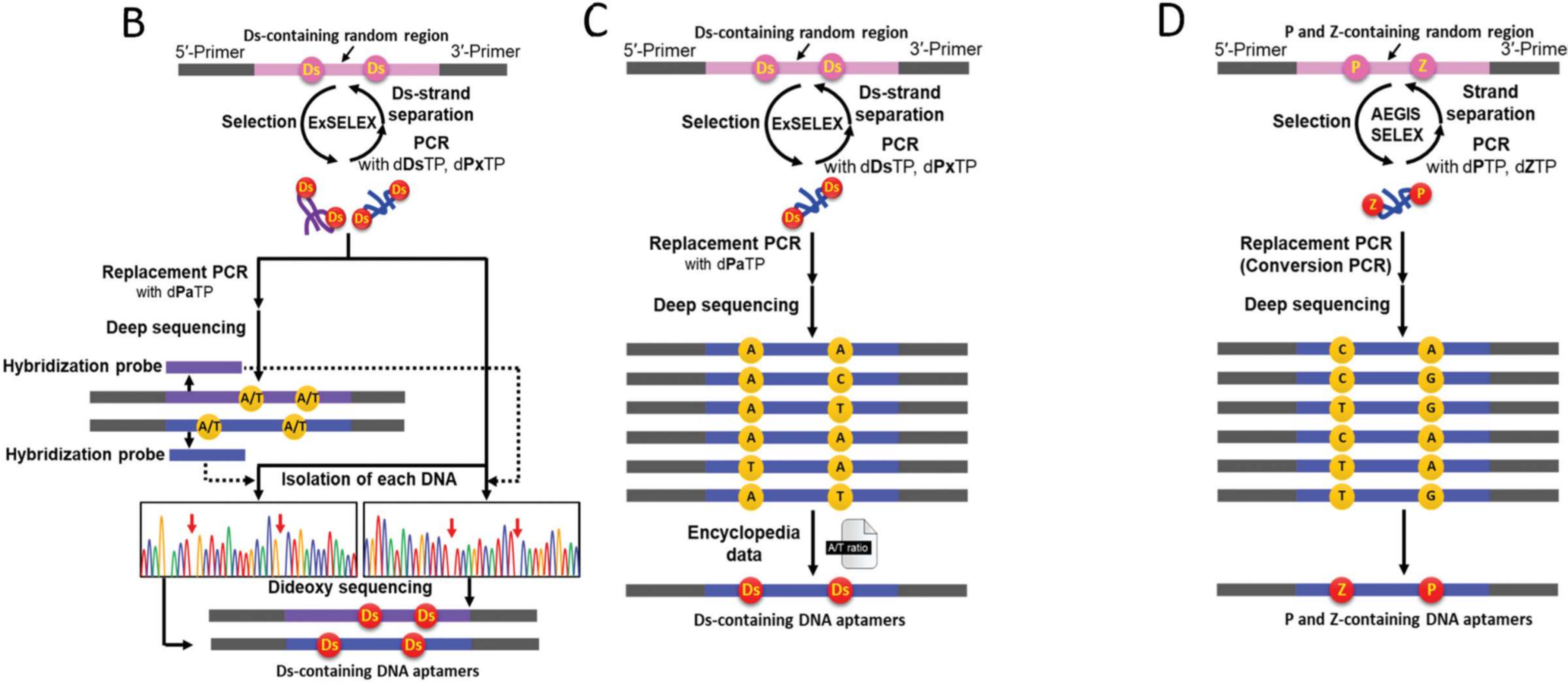
K. Hamashima, I. Hirao *et al. ACS Synth. Biol.* 2019, *8,* 1401-1410

Referring to the encyclopedia data allows for simple and fast determination of the Ds positions.



Comparison of the replacement patterns between two conditions enables the Ds positions to be distinguished from other natural-base positions.

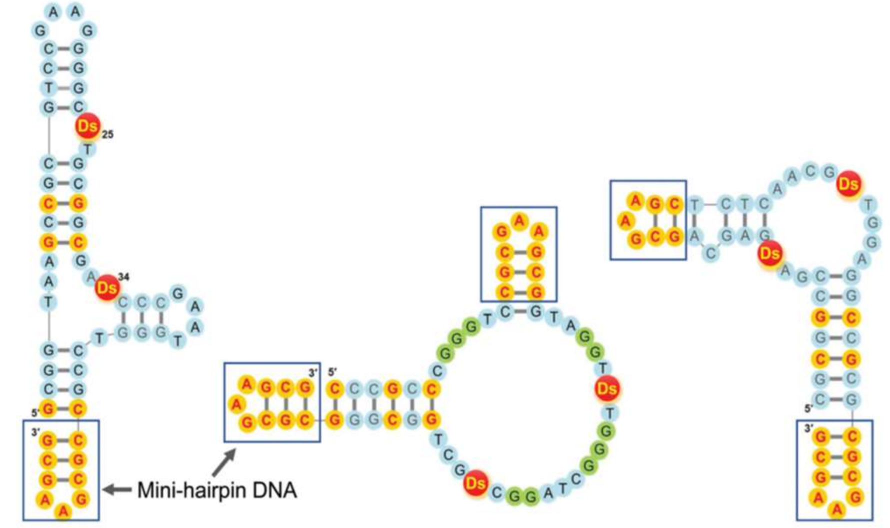
Ds-Pa aptamer selection using deep sequencing



M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

Ds-Pa aptamer selection using deep sequencing

Secondary structures of the optimized Ds-DNA aptamers targeting VEGF165, IFNγ, and vWF.

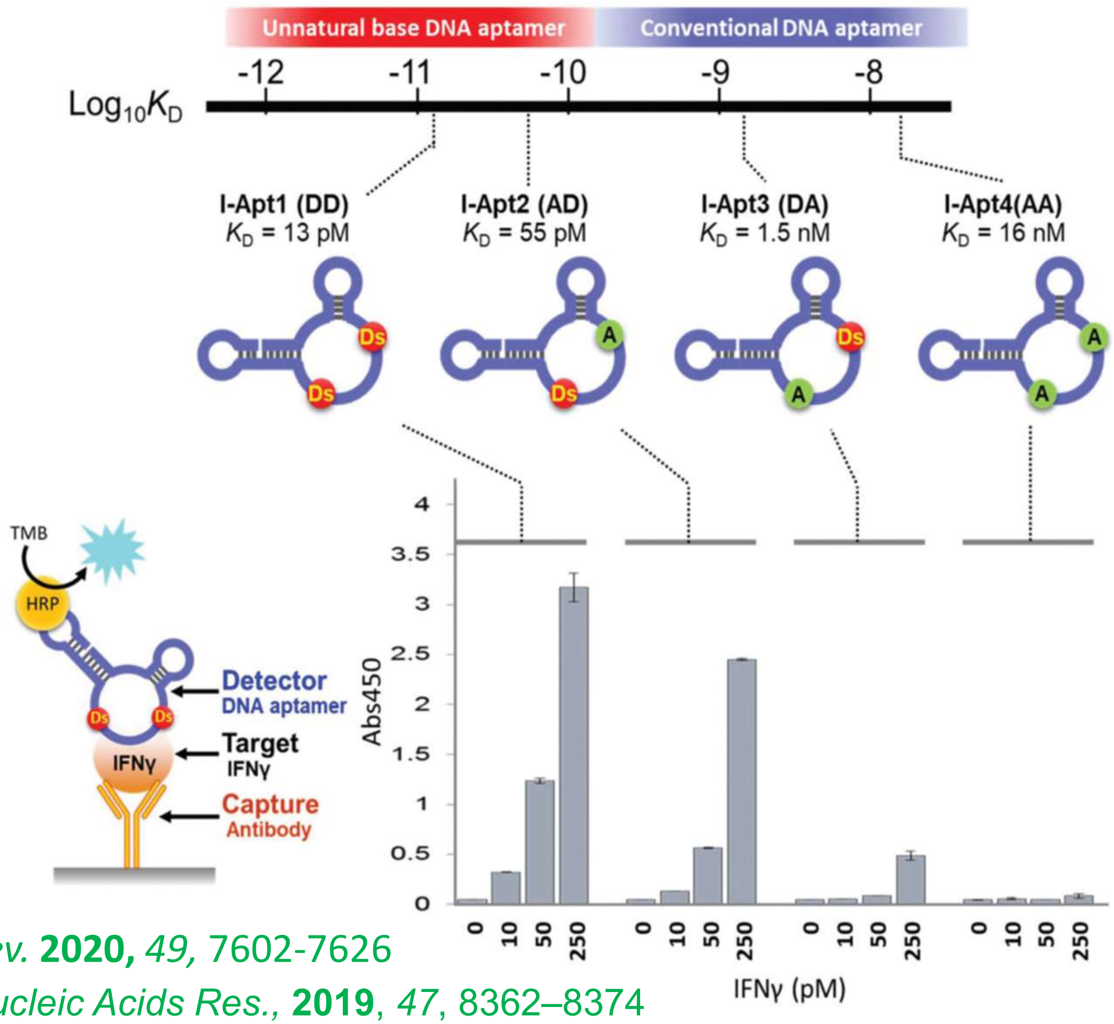


anti-VEGF-aptamer
K_D= ~1 pM
T_m= 80.8°C

anti-IFNγ-aptamer
K_D= 33 pM
T_m= 64.2°C

anti-vWF-aptamer
K_D= 61 pM
T_m= 76.5°C

Generation of a molecular affinity ruler using anti-IFNγ Ds-DNA aptamer variants, replacing the Ds base with the natural A base.

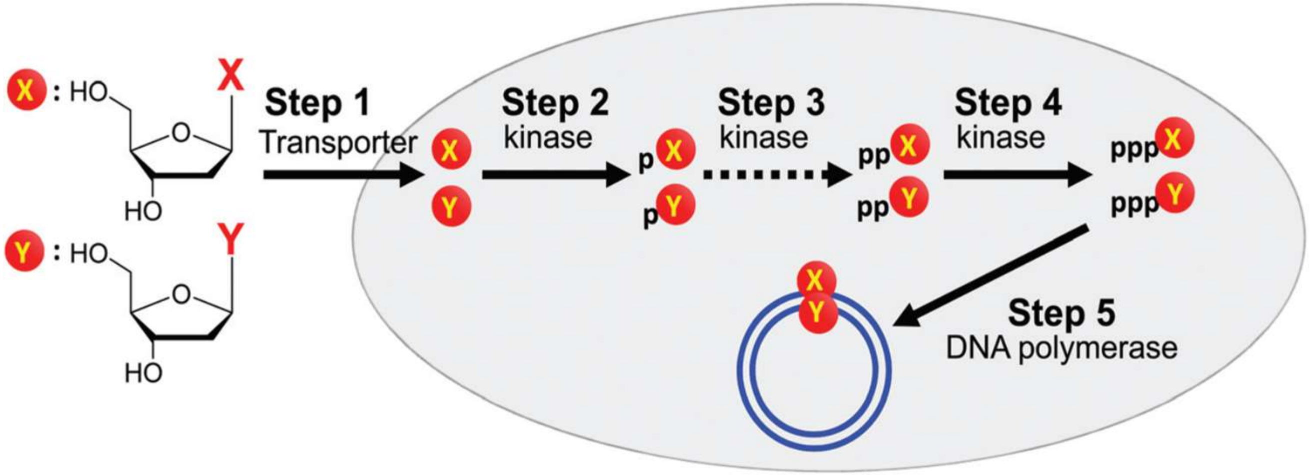


M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

M. Kimoto, Y. W. S. Lim and I. Hirao, *Nucleic Acids Res.*, **2019**, *47*, 8362–8374

A semi-synthetic organism with an expanded genetic alphabet

Nucleoside phosphorylation: a biosynthetic pathway to provide unnatural substrates (dX/dYTP) in a cell, using their unnatural nucleosides (dX/dY) as a source.



In Step 2, the nucleoside kinase from D. melanogaster (DmdNK) was explored for its ability to phosphorylate the unnatural nucleosides by three teams in vitro. Benner's team confirmed that the kinase in Step 4 was available for their Z and P bases in vitro. All UB triphosphates are, at least, accepted by the Klenow fragment of E. coli Pol I in vitro

M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

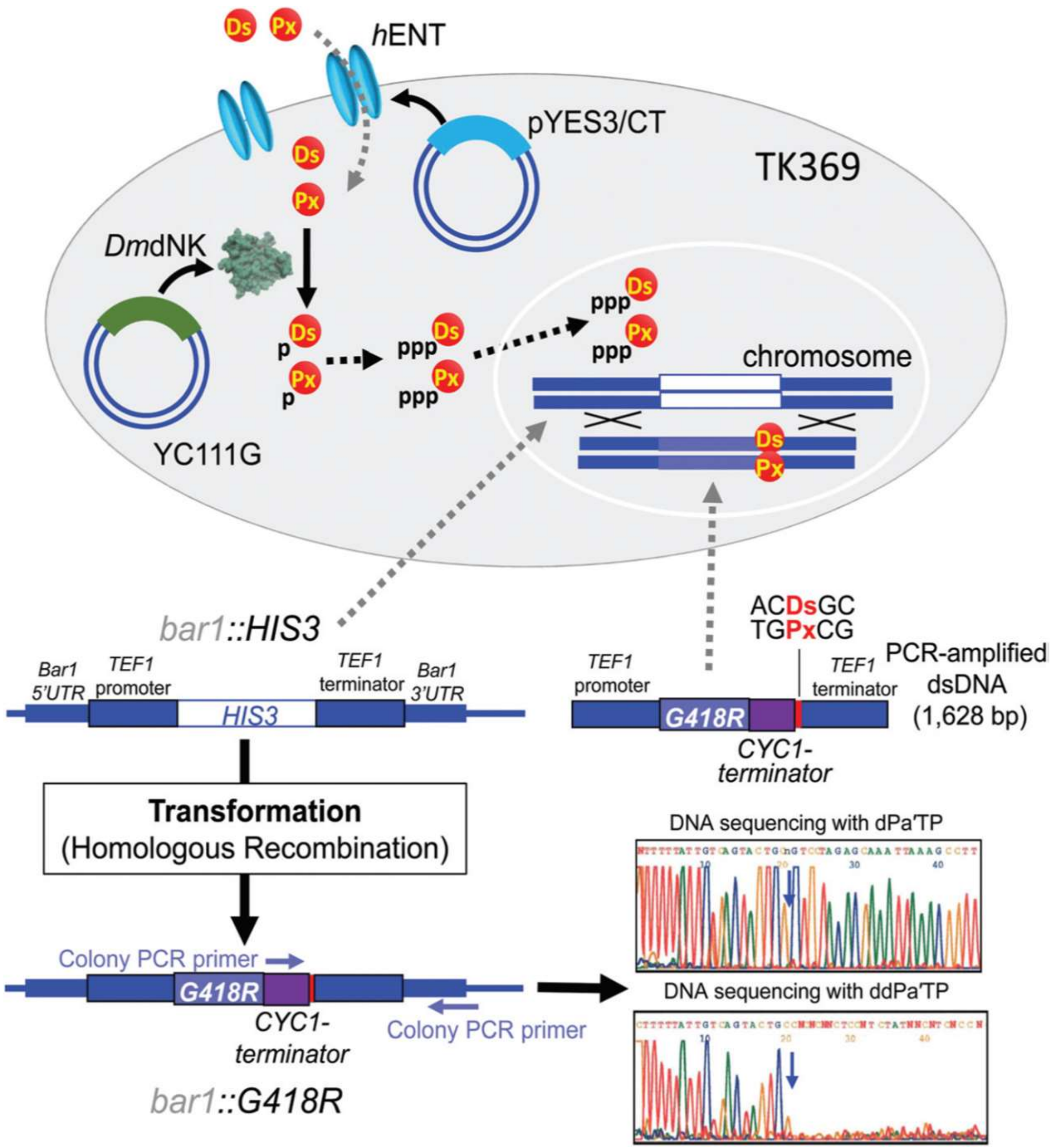
A semi-synthetic organism with an expanded genetic alphabet

Homologous recombination strategy using the yeast strain TK369, to integrate dsDNA with the Ds–Px pair.

DmdNK – nucleoside kinase
hENT - human equilibrium nucleoside transporter

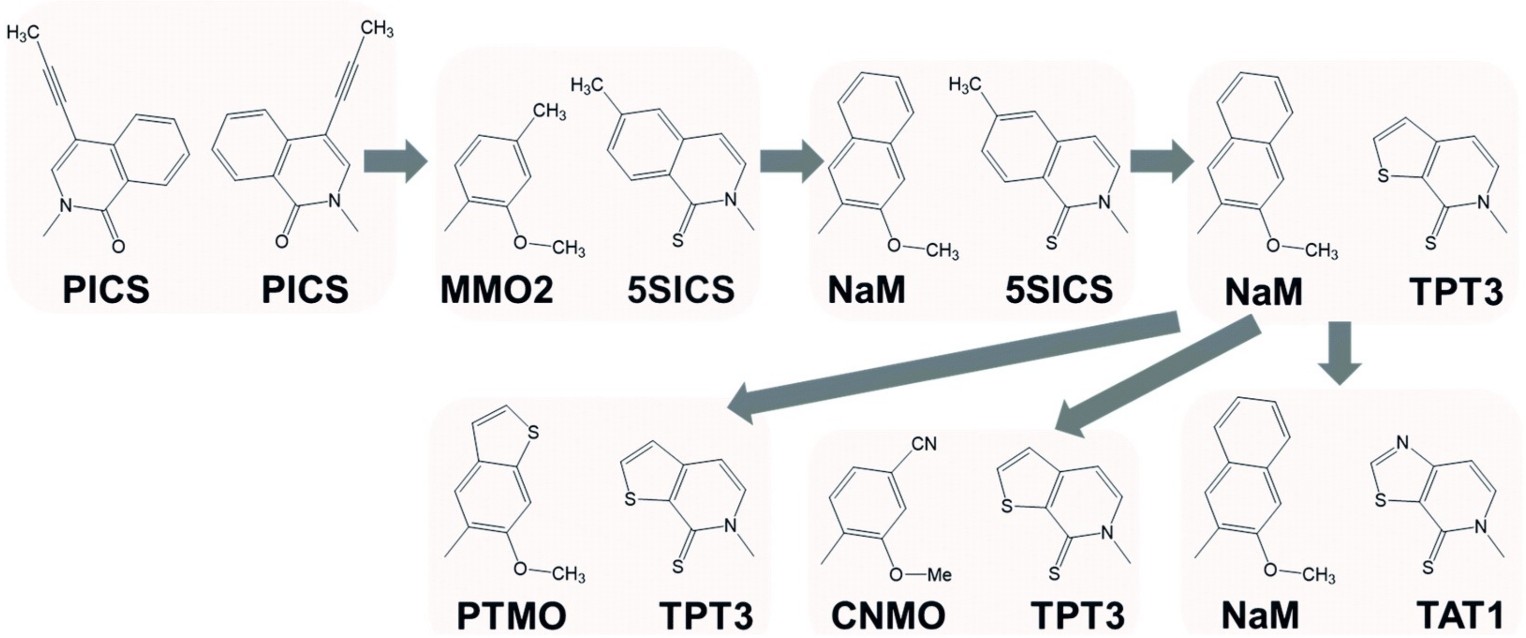
The Ds and Px nucleosides supplied in the media.
The long dsDNA with the Ds–Px pair was prepared by fusion/overlapping PCR,

Analysis: colony PCR with dDsTP and dPxTP, then modified Sanger sequencing.



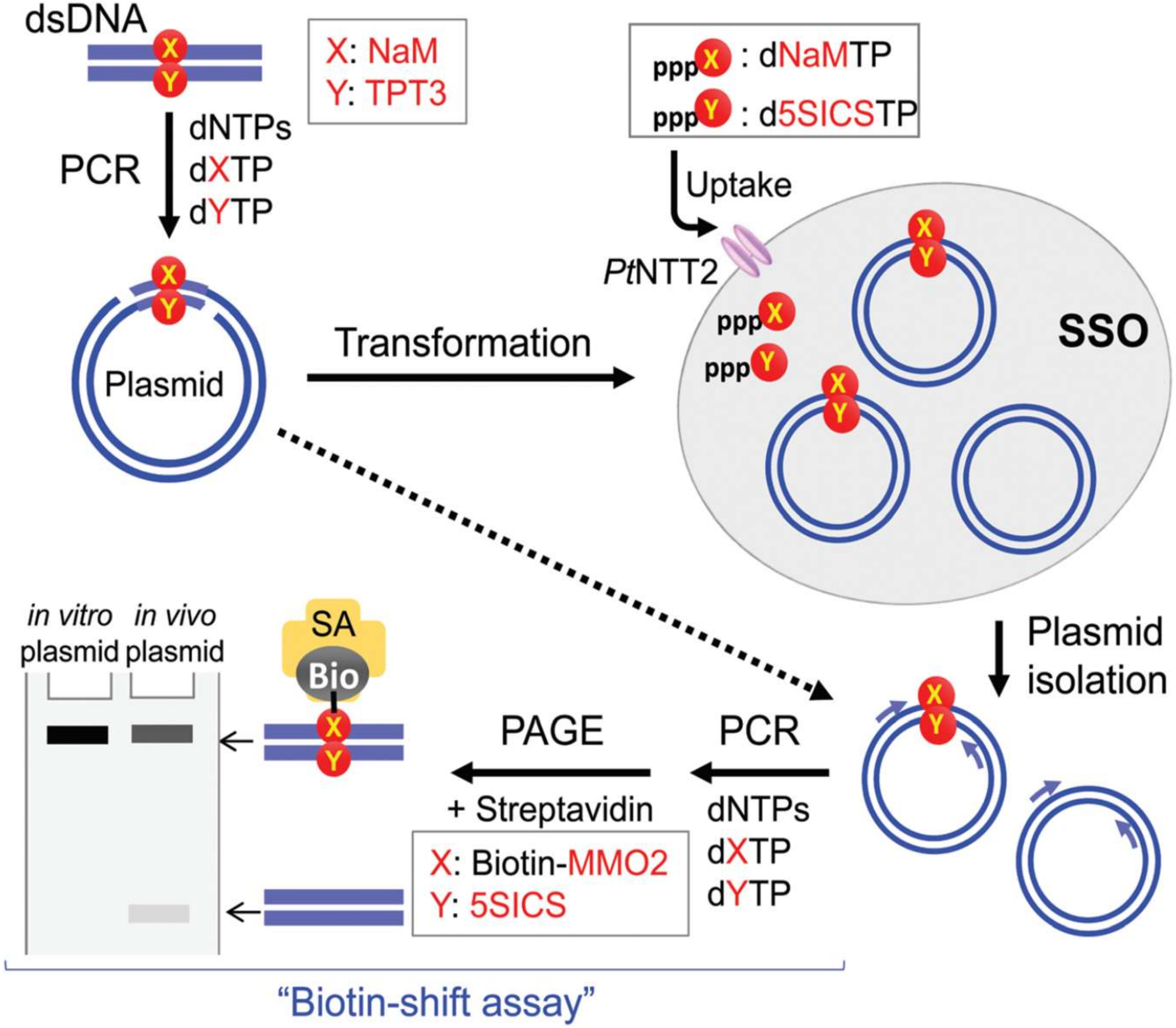
M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

Hydrophobic unnatural base pairs



Romesberg (1999): **PICS–PICS**, but the polymerase reaction is paused at the PICS–PICS pairing position due to self-stacking.
Romesberg (2008): the **MMO2–5SICS** pair - the methyl group of 5SICS prevented the disfavored 5SICS–5SICS self-pairing.
Further optimized into the **NaM–5SICS** pair (>99% fidelity in PCR and transcription).
Romesberg (2014) the **NaM–TPT3** pair (systematic *in vitro* replication screening) ➔ semi-synthetic organism (SSO) of *E. coli* with six-letter DNA.
Optimizations: the **CNMO–TPT3** pair to increase the UBP retention on a plasmid within the SSO,
the **NaM–TAT1** pair for efficient transcription and translation in the SSO.
Using these UBP systems, they reported protein synthesis with the site-specific incorporation of unnatural amino acids into proteins in the SSO.

A semi-synthetic organism with an expanded genetic alphabet



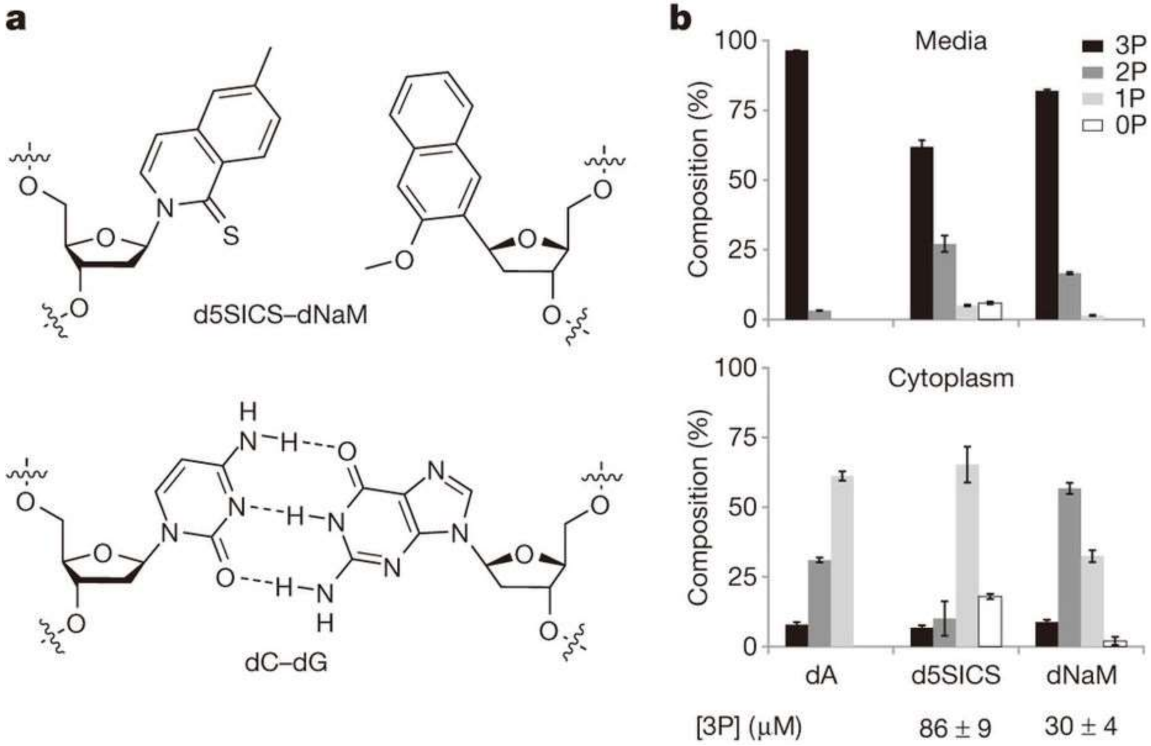
The first creation of an SSO, using the NaM–TPT3 and NaM–5SCIS pairs and E. coli C41(DE3) as the host strain.

The plasmid containing the NaM–TPT3 pair was prepared by PCR and transformed into E. coli C41(DE3) with a PtNTT2 overexpression system. The transformants were cultured in inorganic phosphate-rich growth media in the presence of the NaM and 5SCIS triphosphates.

The propagated plasmids were isolated and subjected to the analysis of the NaM–5SCIS pair retention by a biotin-shift assay

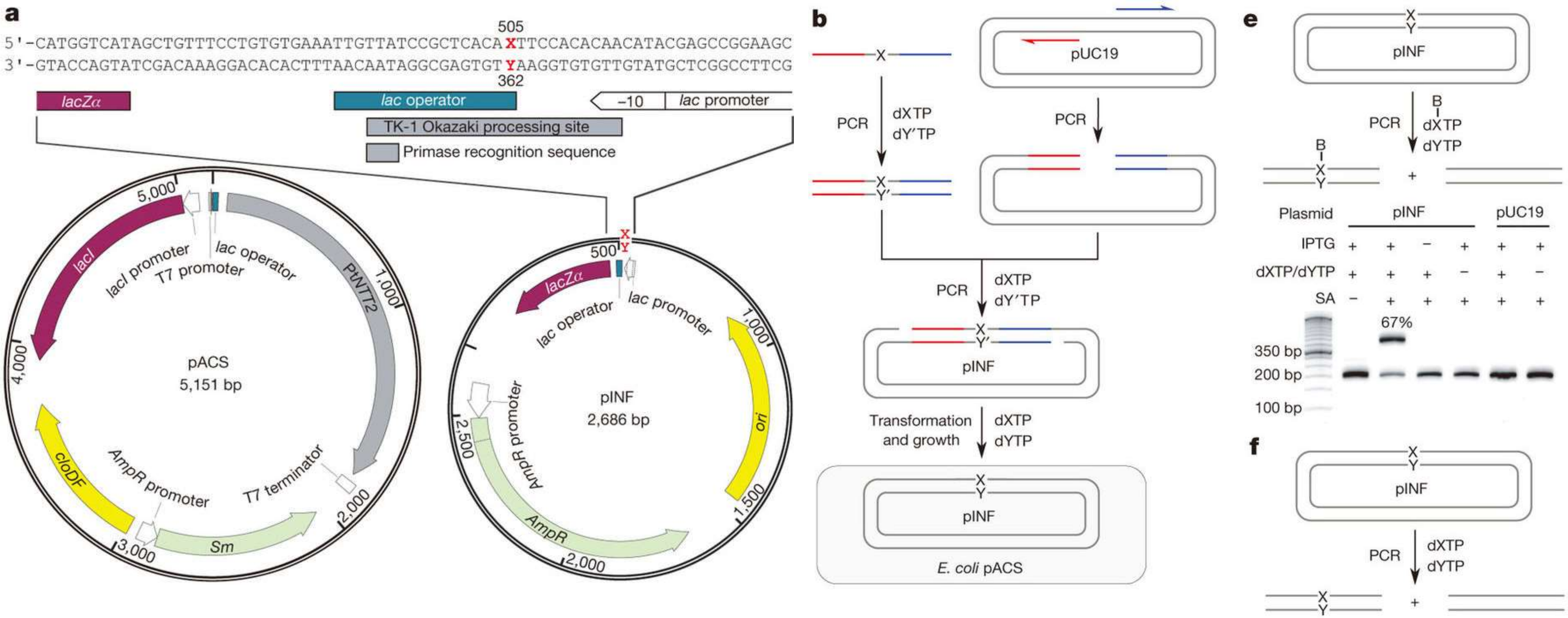
Malyshev, Denis A.; Romesberg, Floyd E. *et al.* *Nature* **2014**, *509,* 385–388

A semi-synthetic organism with an expanded genetic alphabet



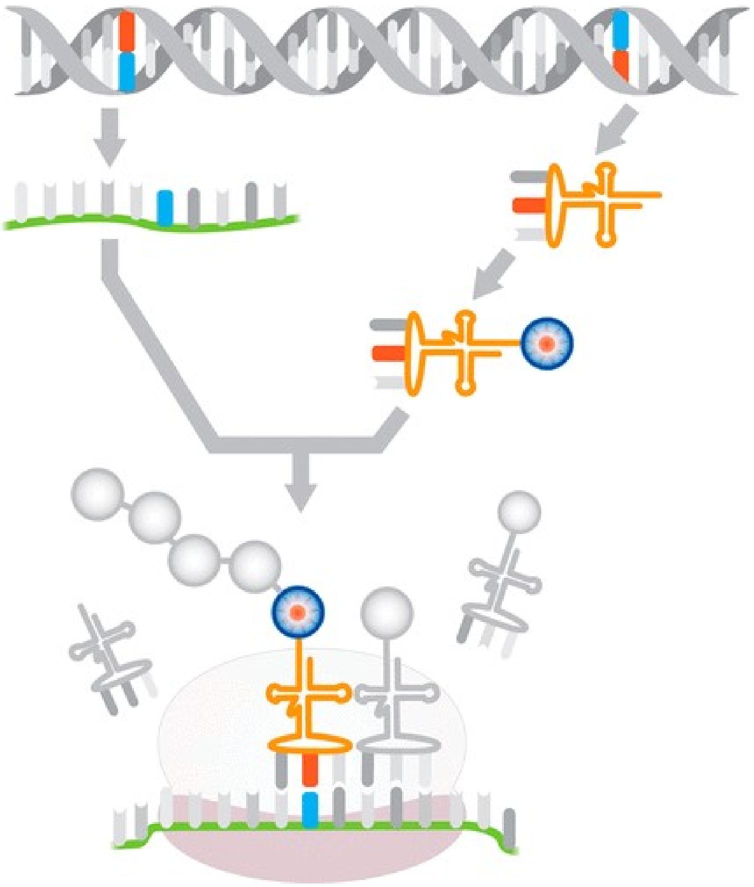
a, Chemical structure of the d5SICS–dNaM UBP compared to the natural dG–dC base pair. **b**, Composition analysis of d5SICS and dNaM in the media (top) and cytoplasmic (bottom) fractions of cells expressing *Pt*NTT2 after 30 min incubation; dA shown for comparison. 3P, 2P, 1P and 0P correspond to triphosphate, diphosphate, monophosphate and nucleoside, respectively; [3P] is the intracellular concentration of triphosphate.

Malyshev, Denis A.; Romesberg, Floyd E. *et al.* *Nature* **2014**, *509,* 385–388



Malyshev, Denis A.; Romesberg, Floyd E. *et al.* *Nature* **2014**, *509,* 385–388

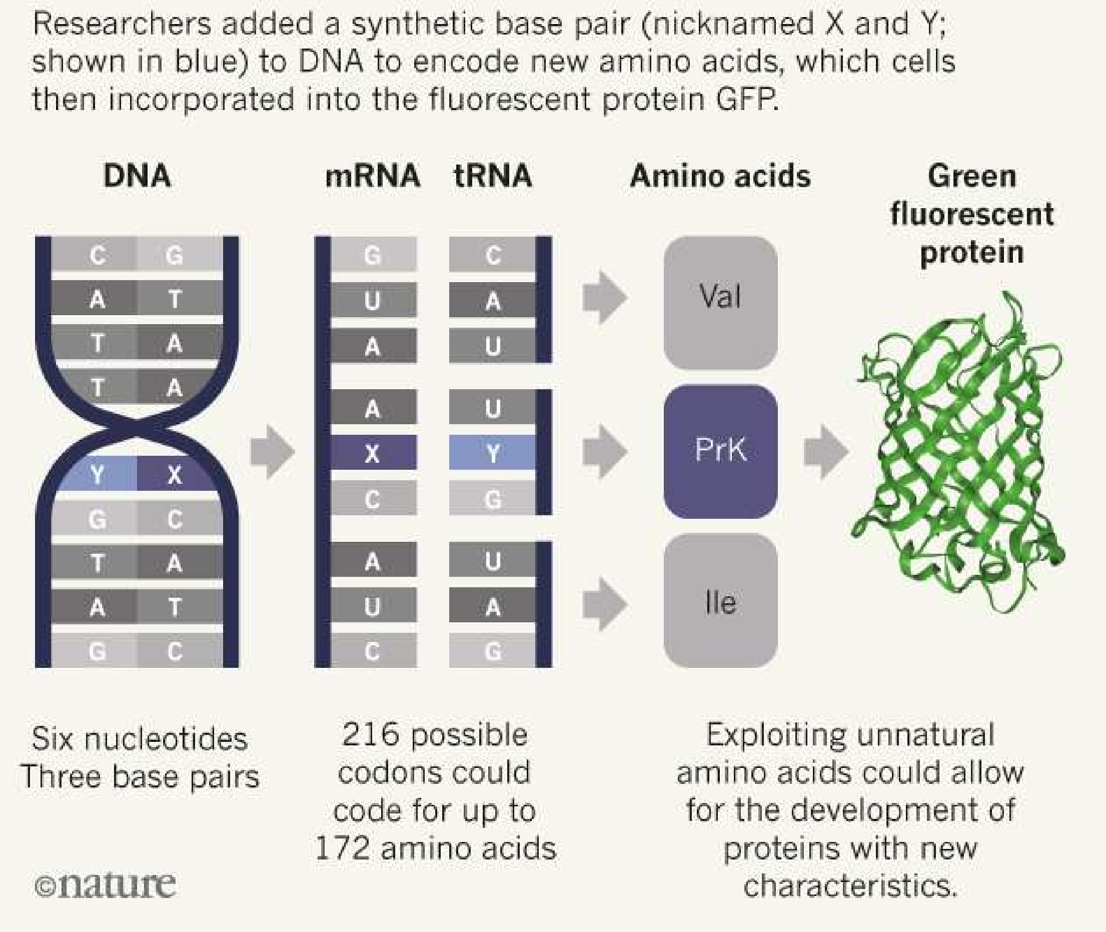
A semi-synthetic organism with an expanded genetic alphabet



The SSO replicates DNA containing the d**NaM**-d**TPT3** UBP (blue and red, respectively), transcribes mRNA and tRNA with complementary codons and anticodons containing **NaM** or **TPT3**, uses an orthogonal synthetase to charge the tRNA with an ncAA, and uses the charged tRNA to translate the mRNA into proteins containing ncAAs

Y. Zhang, F. Romesberg, *Biochemistry* **2018**, *57, 15*, 2177–2178

A semi-synthetic organism with an expanded genetic alphabet



Researchers added a synthetic base pair (nicknamed X and Y; shown in blue) to DNA to encode new amino acids, which cells then incorporated into the fluorescent protein GFP.

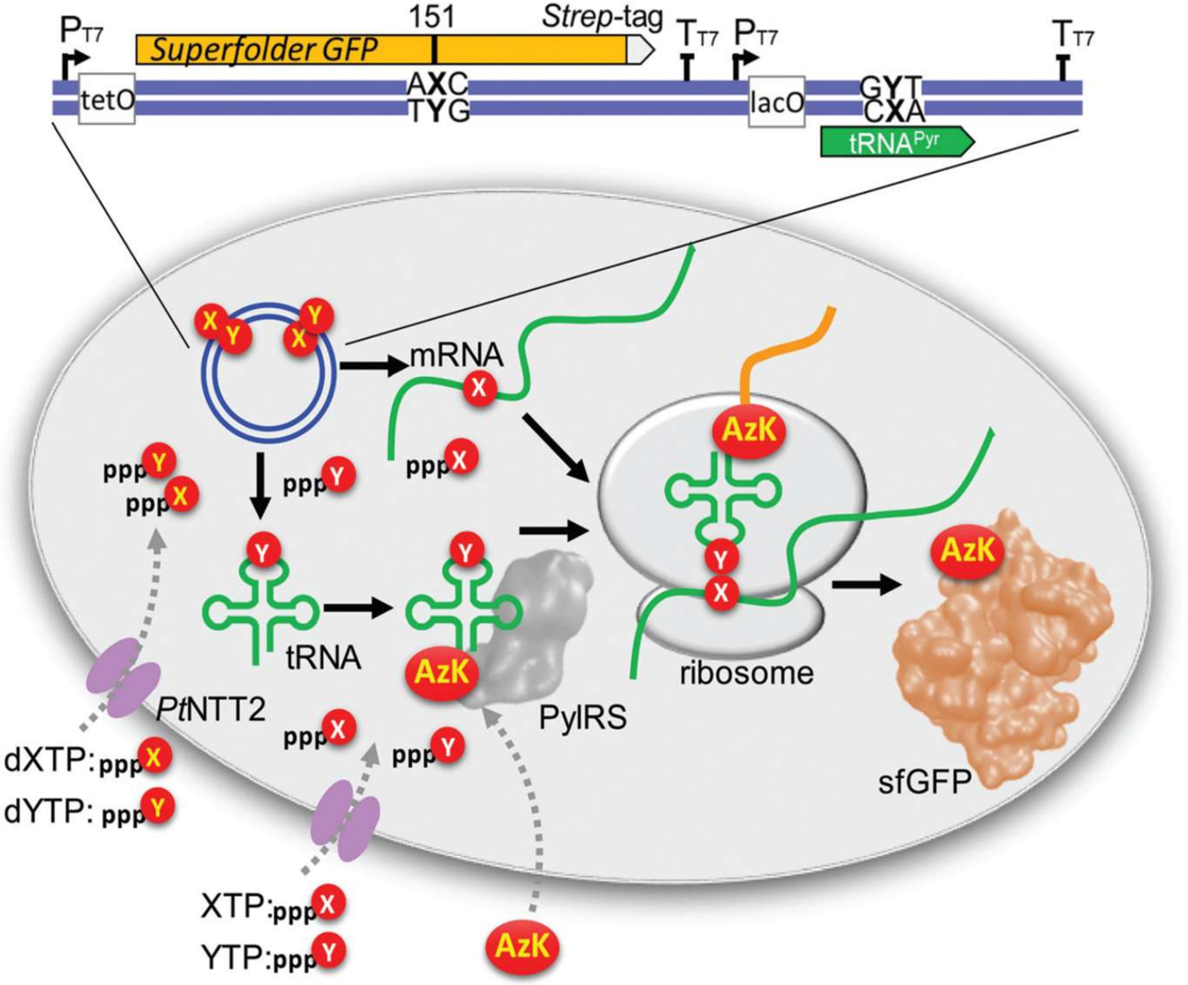
Six nucleotides
Three base pairs

216 possible codons could code for up to 172 amino acids

Exploiting unnatural amino acids could allow for the development of proteins with new characteristics.

Zhang, Y.; Romesberg, Floyd E. *et al.* *Nature* **2017**, *551,* 644-647

A semi-synthetic organism with an expanded genetic alphabet

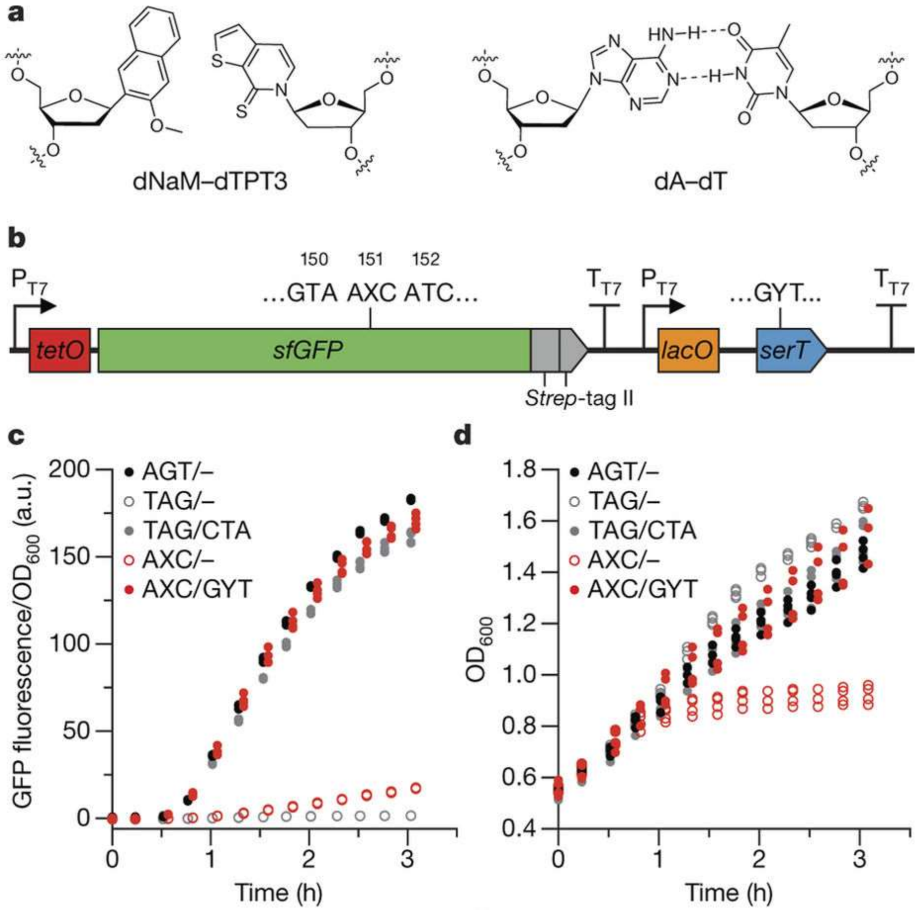


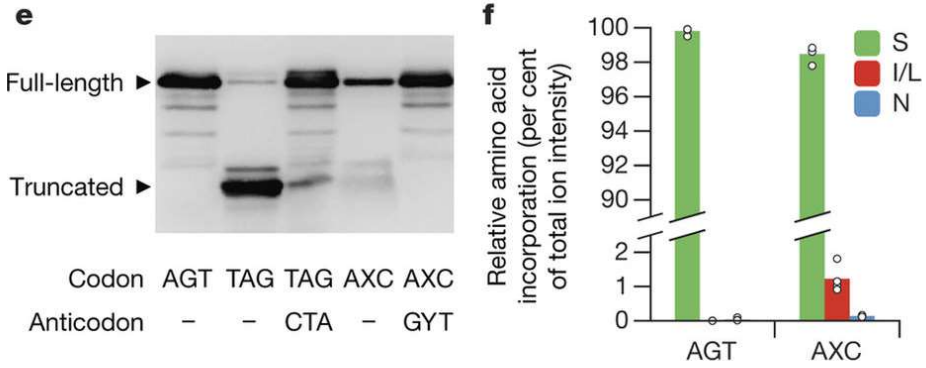
Zhang, Y.; Romesberg, Floyd E. *et al.* *Nature* **2017**, *551,* 644-647

M. Kimoto, I. Hirao *Chem. Soc. Rev.* **2020,** *49,* 7602-7626

A semi-synthetic organism with an expanded genetic alphabet

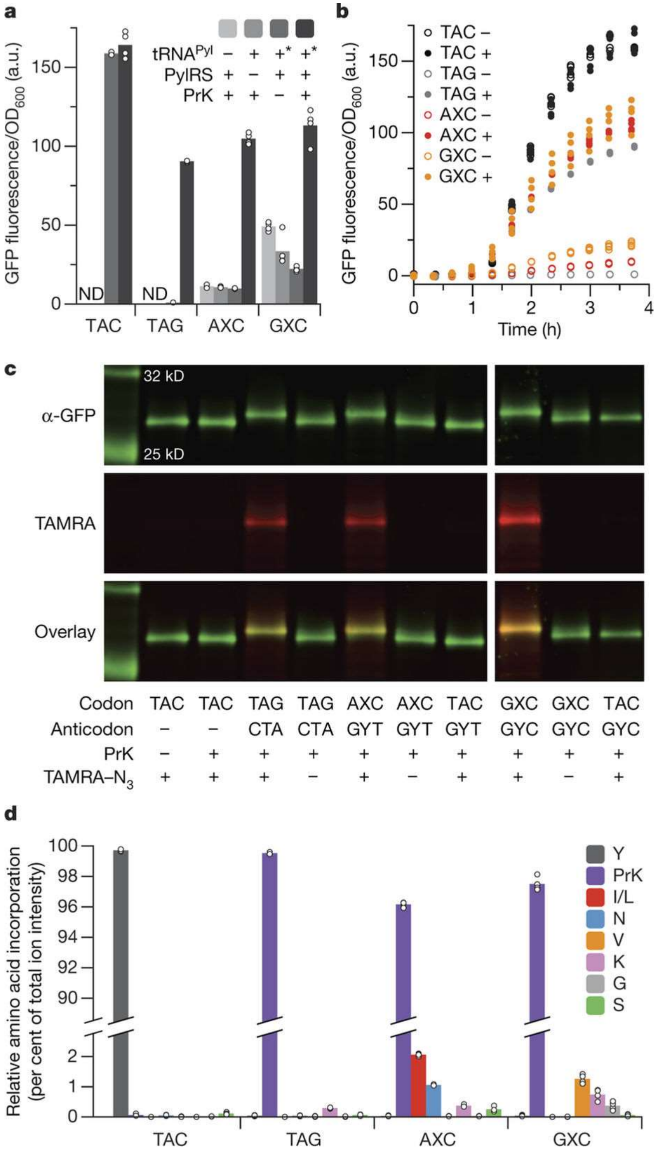
The *in vivo* transcription of DNA containing dNaM and dTPT3 into mRNAs with two different unnatural codons and tRNAs with cognate unnatural anticodons, and their efficient decoding at the ribosome to direct the site-specific incorporation of natural or non-canonical amino acids into superfolder green fluorescent protein. The resulting semi-synthetic organism both encodes and retrieves increased information and should serve as a platform for the creation of new life forms and functions.

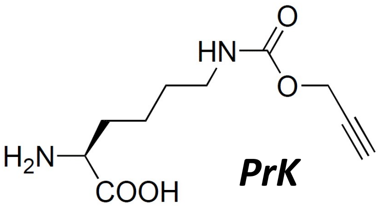




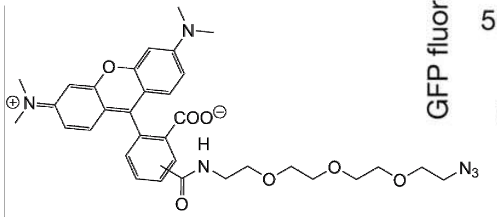
Zhang, Y.; Romesberg, Floyd E. *et al.* *Nature* **2017**, *551,* 644-647

A semi-synthetic organism with an expanded genetic alphabet

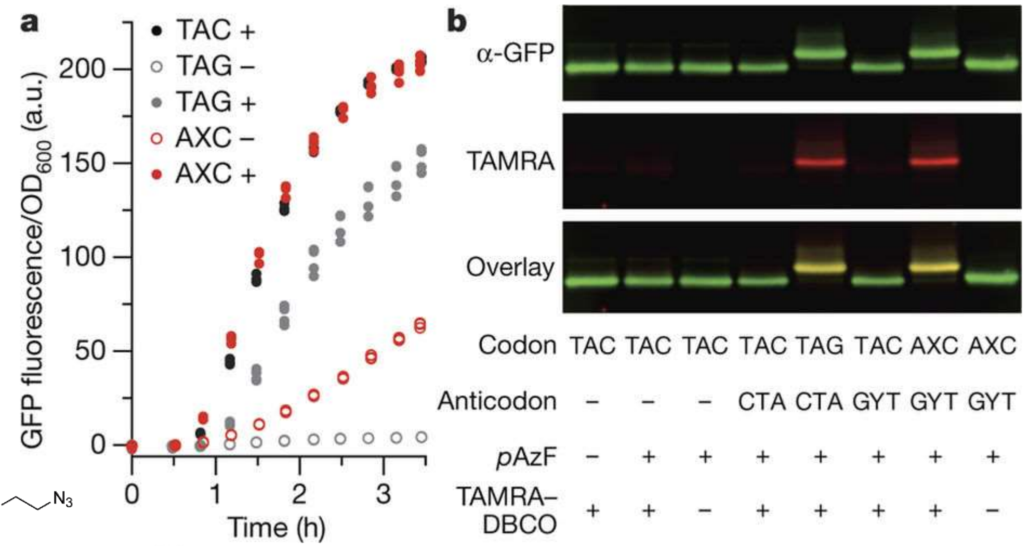


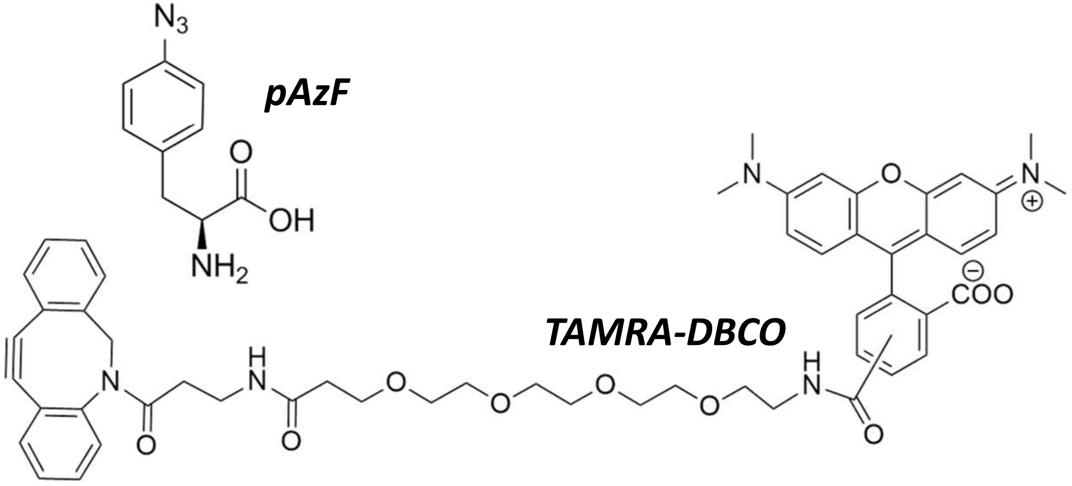


PrK



TAMRA-N$_3$





pAzF

TAMRA-DBCO

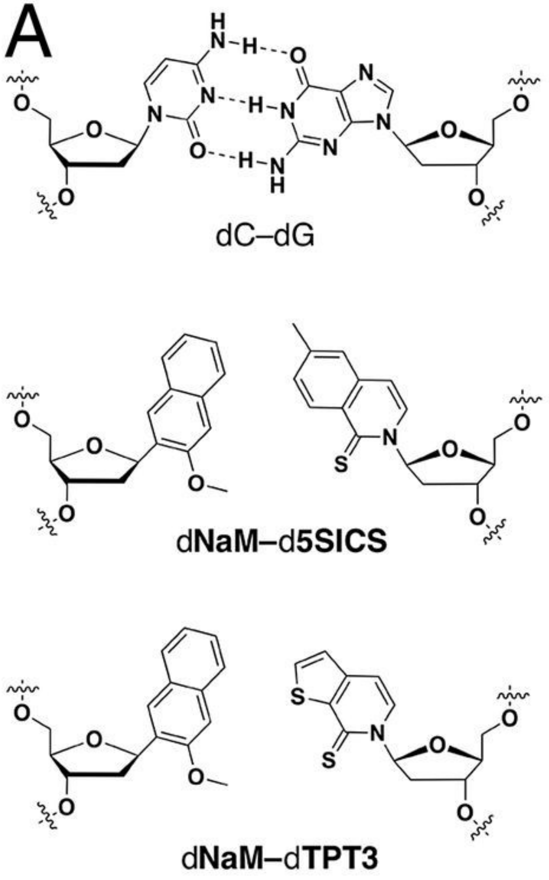
Zhang, Y.; Romesberg, Floyd E. *et al.* *Nature* **2017**, *551,* 644-647

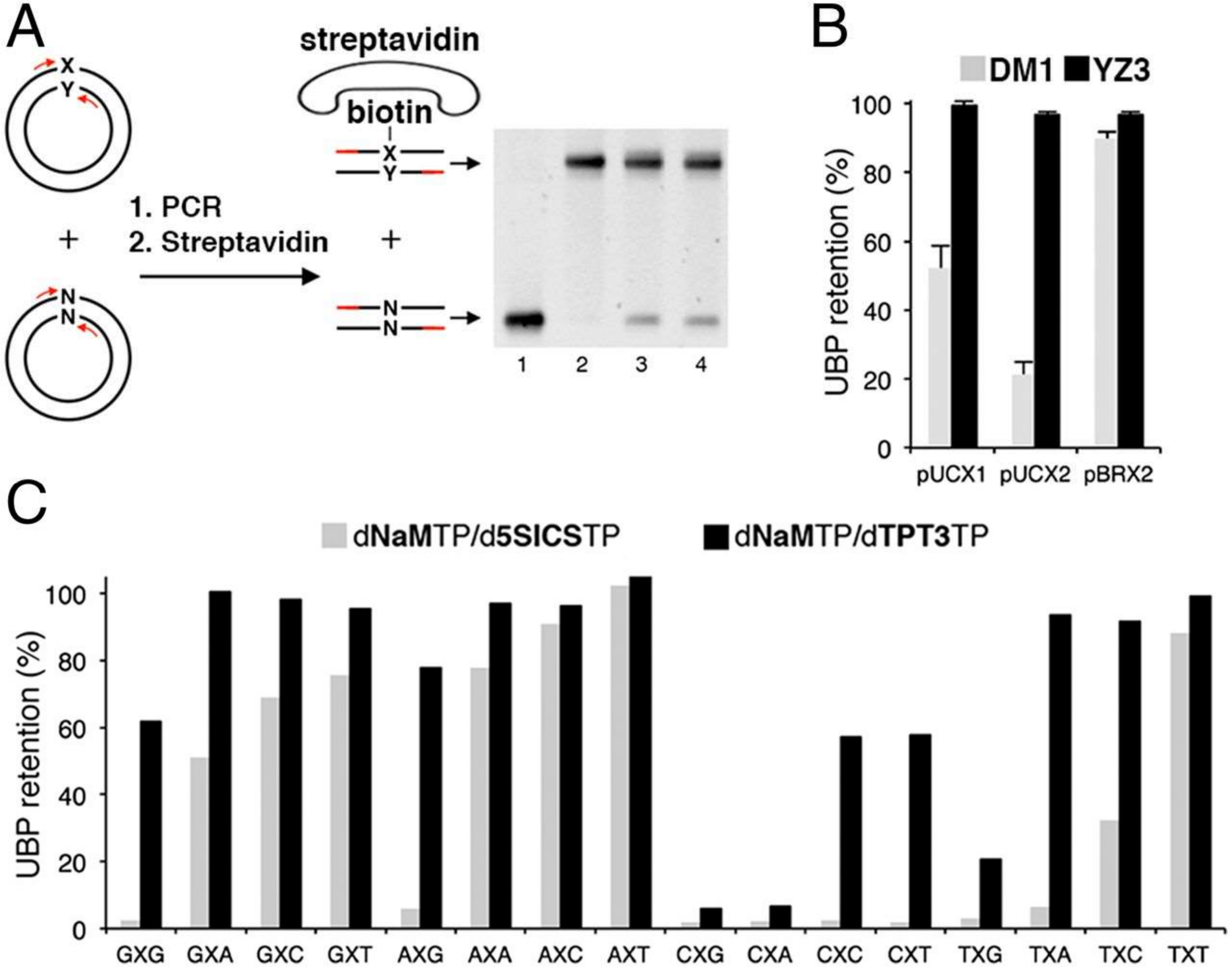
A semi-synthetic organism with an expanded genetic alphabet

- An unnatural base pair (UBP) would increase the information storage potential of DNA
- and semisynthetic organisms (SSOs) that stably harbor this expanded alphabet would thereby have the potential to store and retrieve increased information,
- *Escherichia coli* grown in the presence of the unnatural nucleoside triphosphates dNaMTP and d5SICSTP, and provided with the means to import them via expression of a plasmid-borne nucleoside triphosphate transporter, replicates DNA containing a single dNaM-d5SICS UBP,
- to fortify and vivify the nascent SSO, a more chemically optimized UBP dTPT3 was used, and the power of the bacterial immune response was harnessed by using Cas9 to eliminate DNA that had lost the UBP.
- The optimized SSO grows robustly, constitutively imports the unnatural triphosphates, and is able to indefinitely retain multiple UBPs in virtually any sequence context. This SSO is thus a form of life that can stably store genetic information using a six-letter, three-base-pair alphabet

Romesberg, Floyd E. *et al.* *PNAS* **2017**, *114,* 1317-1322

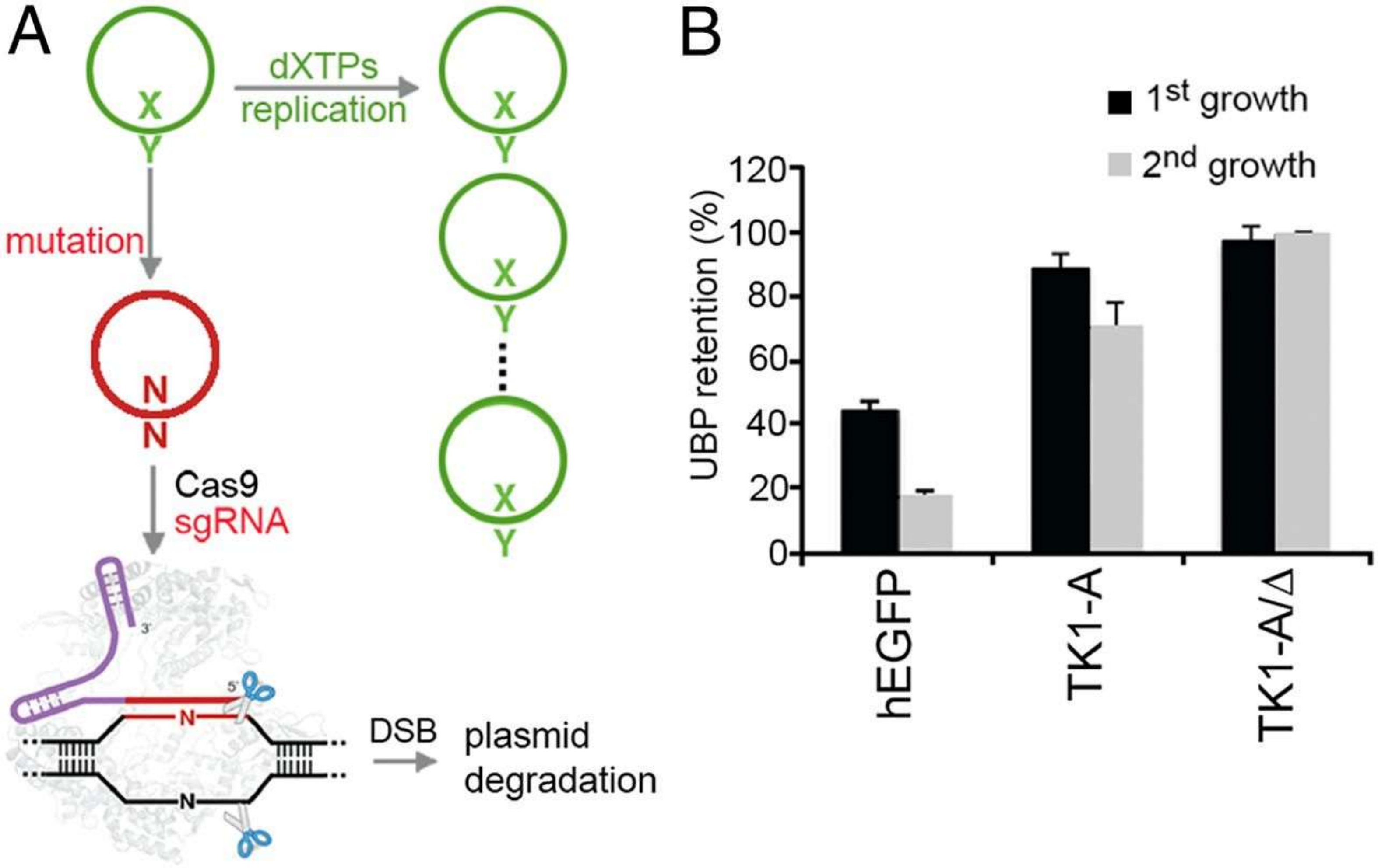
A semi-synthetic organism with an expanded genetic alphabet





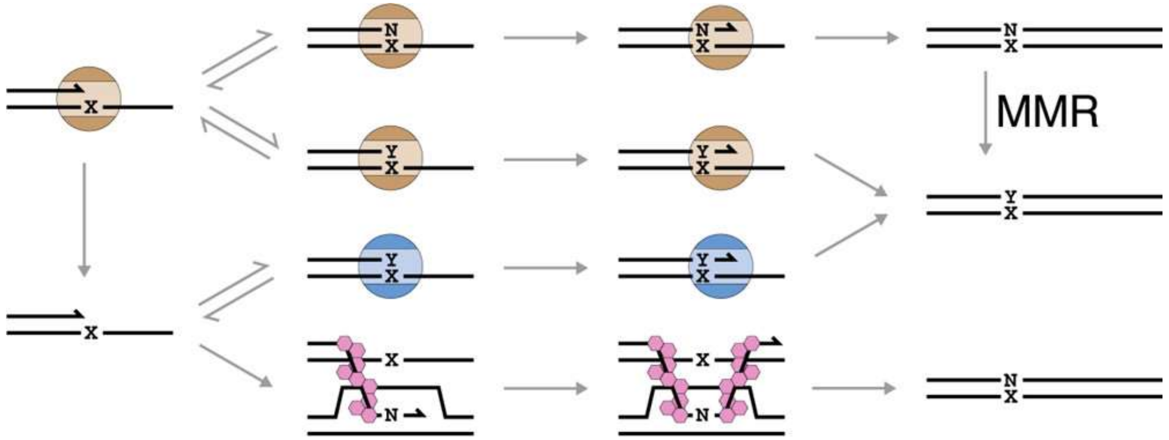
Romesberg, Floyd E. *et al.* *PNAS* **2017**, *114,* 1317-1322

A semi-synthetic organism with an expanded genetic alphabet



Romesberg, Floyd E. *et al.* *PNAS* **2017**, *114,* 1317-1322

Reprograming the replisome of a semi-synthetic organism

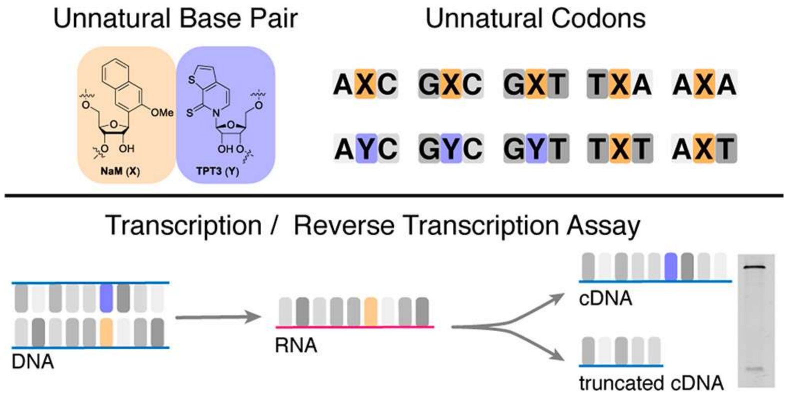


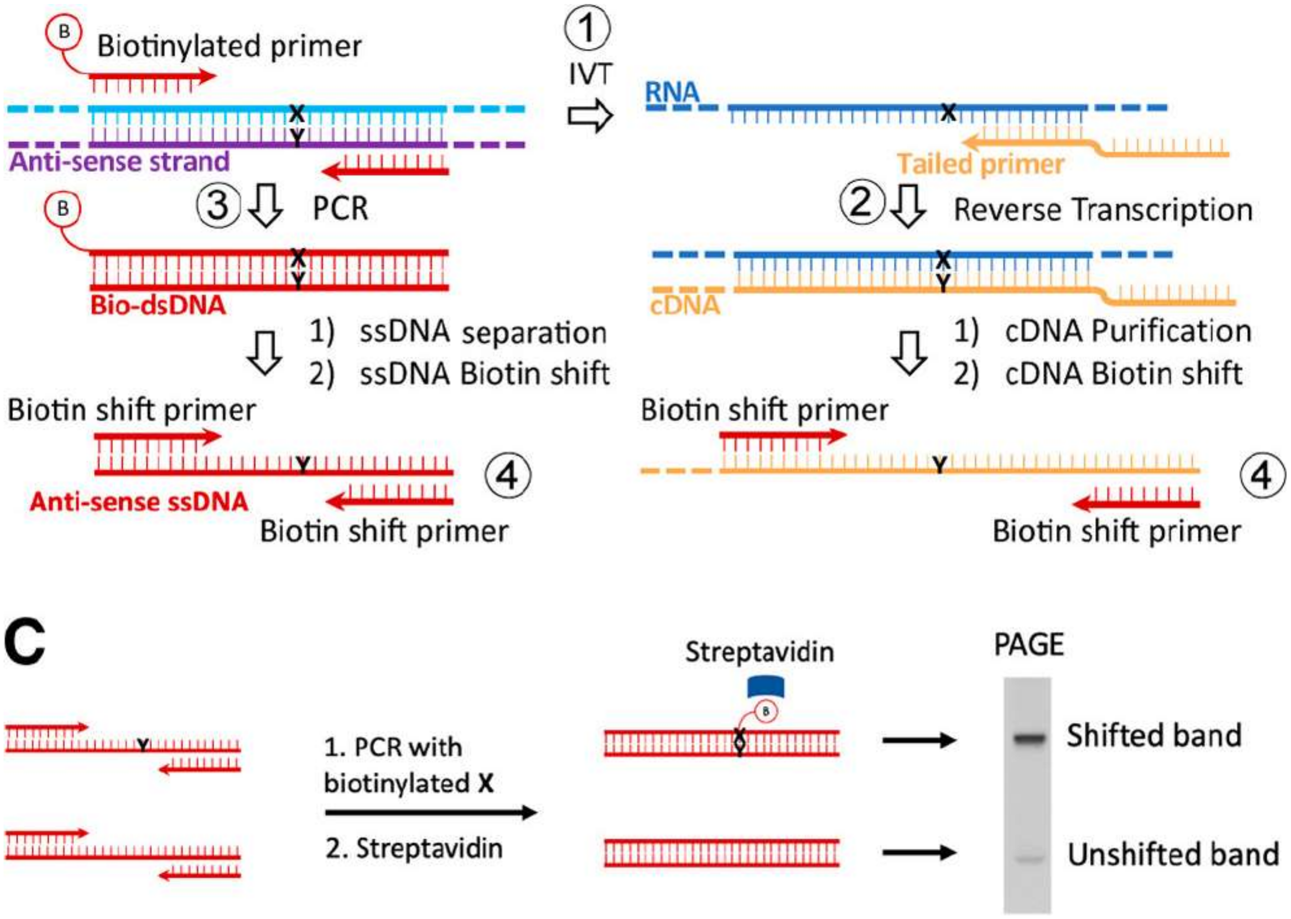
Proposed mechanism of UBP replication. *UBP retention is mediated by the activities of Pol III (tan), Pol II (blue), and MMR. UBP loss is mediated by polymerase replication errors or RecA (pink)-mediated RER.*

UBPs are not well retained in all sequences, limiting the information that can be encoded, and are invariably lost upon extended growth. The contributions of the *E. coli* DNA replication and repair machinery to the propagation of DNA containing d**NaM**-d**TPT3** have been explored and showed that replication by DNA polymerase III, supplemented with the activity of polymerase II and methyl-directed mismatch repair contribute to retention of the UBP and that recombinational repair of stalled forks is responsible for the majority of its loss.

M. P. Ledbetter, R. J. Karadeema, and F. E. Romesberg *J. Am. Chem. Soc.* **2018**, *140, 2,* 758-765

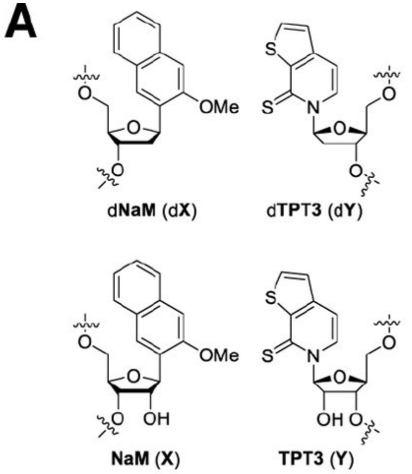
Transcription and reverse transcription in vitro and in SSO

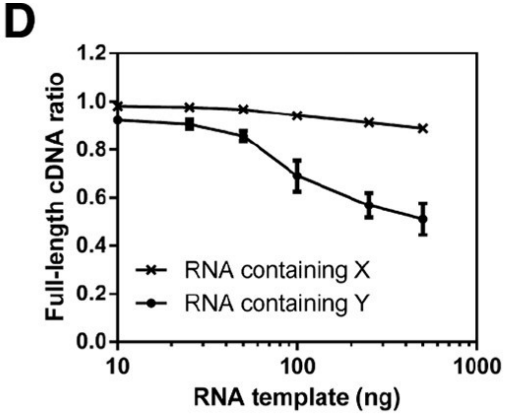


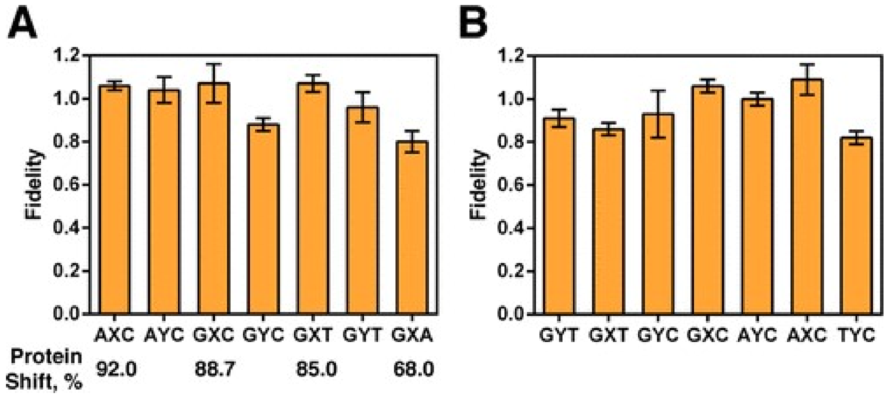


A. X.-Z. Zhou, X. Dong, F. Romesberg,
J. Am. Chem. Soc., **2020**, *142*, 19029-19032

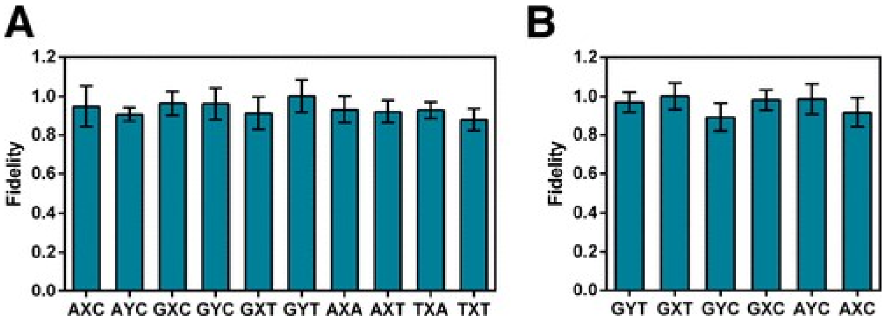
Transcription and reverse transcription in vitro and in SSO







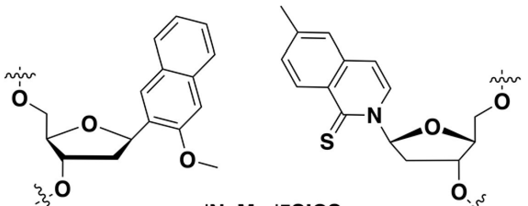
Transcription fidelity measured by T-RT assay of (A) mRNA and (B) tRNA extracted from SSO in vivo translation experiments



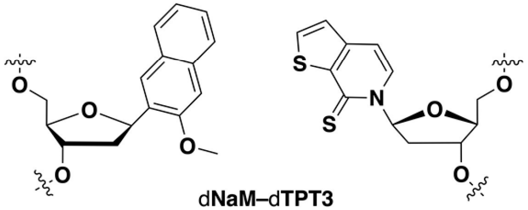
Transcription fidelity measured by T-RT assay of (A) mRNA and (B) tRNA made by in vitro transcription using T7 RNA polymerase.

A. X.-Z. Zhou, X. Dong, F. Romesberg,
J. Am. Chem. Soc., **2020**, *142*, 19029-19032

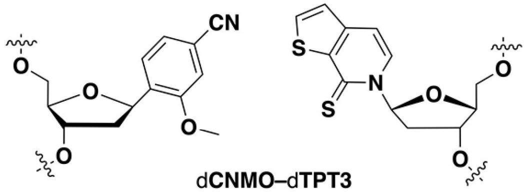
Expanded genetic alphabet - 2019



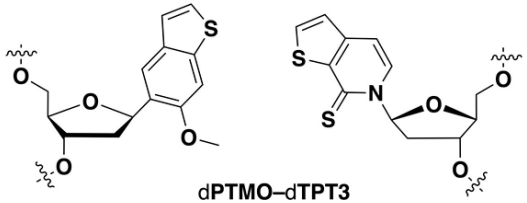
dNaM–d5SICS



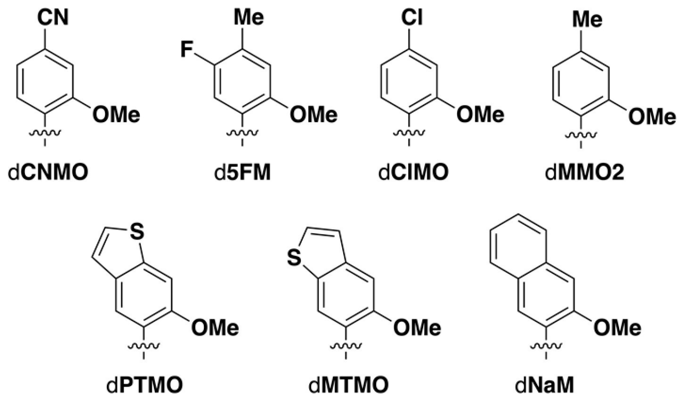
dNaM–dTPT3



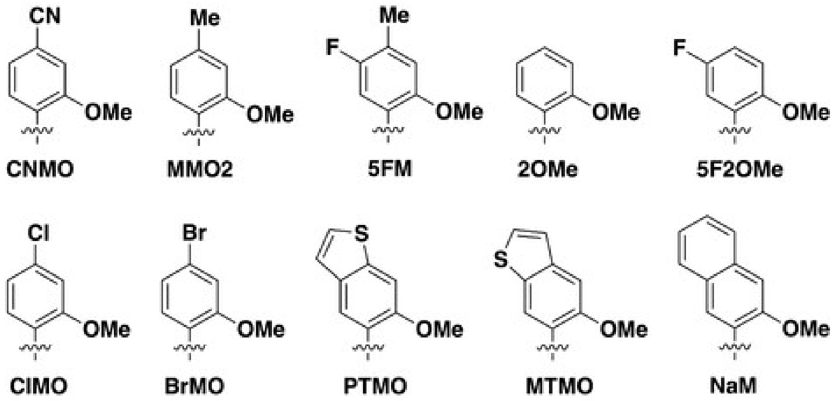
dCNMO–dTPT3



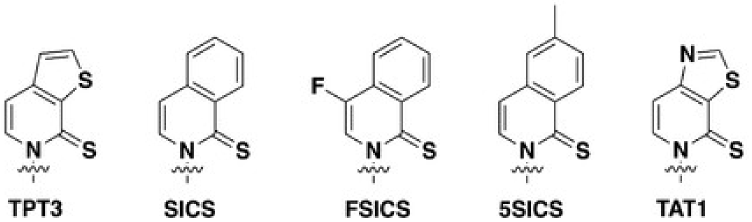
dPTMO–dTPT3



dXTP analogues



XTP analogues



YTP analogues

Romesberg, Floyd E. *et al.*
J. Am. Chem. Soc. **2019**, *141, 27,* 10644–10653

Unnatural nucleobases - overview