



- (51) **International Patent Classification:**
C12Q 1/68 (2006.01) *C07H 21/04* (2006.01)
- (21) **International Application Number:**
PCT/US2014/033793
- (22) **International Filing Date:**
11 April 2014 (11.04.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
61/853,758 11 April 2013 (11.04.2013) US
61/854,138 18 April 2013 (18.04.2013) US
- (71) **Applicant:** CARNEGIE MELLON UNIVERSITY
[US/US]; 5000 Forbes Avenue, Pittsburgh, Pennsylvania
15213 (US).
- (72) **Inventors:** LY, Danith, H.; Pittsburgh, Pennsylvania (US).
GOPALSAMY Suresh, Kumar; 34 Hussain Subedar
Street, Alandur, Chennai 600016, Tamilnadu, India (IN).
MANNA, Arunava; Pittsburgh, Pennsylvania (US).
- (74) **Agents:** HIRSHMAN, Jesse, A. et al.; The Webb Law
Firm, One Gateway Center, 420 Fort Duquesne Boulevard,
Suite 1200, Pittsburgh, Pennsylvania 15222 (US).

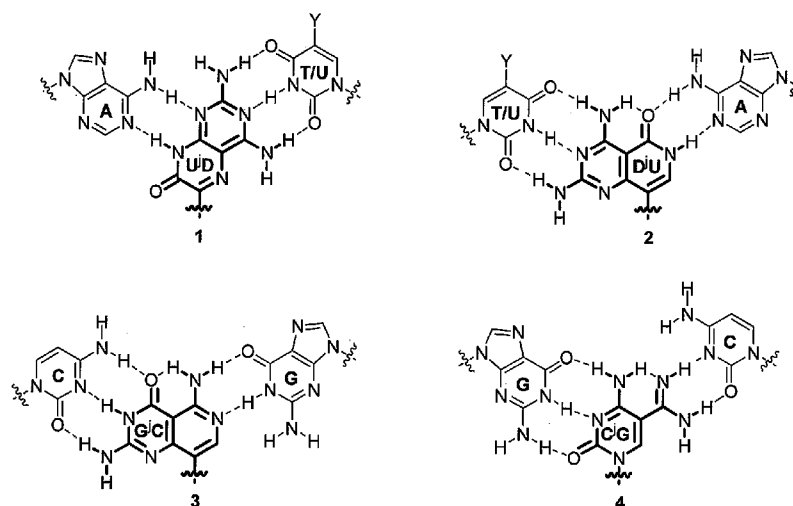
(81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

(54) **Title:** DIVALENT NUCLEOBASE COMPOUNDS AND USES THEREFOR

**Fig. 1B**

(57) **Abstract:** Described herein are novel divalent nucleobases that each bind two nucleic acid strands, matched or mismatched when incorporated into a nucleic acid or nucleic acid analog backbone (a genetic recognition reagent, or genetic recognition reagent). In one embodiment, the genetic recognition reagent is a peptide nucleic acid (PNA) or gamma PNA (?PNA) oligomer. Uses of the divalent nucleobases and monomers and genetic recognition reagents containing the divalent nucleobases also are provided.

DIVALENT NUCLEOBASE COMPOUNDS AND USES THEREFOR

STATEMENT REGARDING FEDERAL FUNDING

[0001] This invention was made with government support under the National Science Foundation CHE-1012467. The government has certain rights in this invention.

CROSS REFERENCE TO RELATED APPLICATIONS

[0002] This application claims the benefit of copending U.S. Provisional Application, Nos. 61/853,758, filed April 11, 2013, and 61/854,138, filed April 18, 2013, each of which is incorporated herein by reference in its entirety.

BACKGROUND

1. Field of the Invention

[0003] Described herein are nucleobases, polymer monomers comprising the nucleobases and nucleic acids and analogs thereof comprising the nucleobases. Also described herein are methods of use of the nucleobases, polymer monomers comprising the nucleobases and nucleic acids and analogs thereof comprising the nucleobases.

2. Description of the Related Art

[0004] For most organisms, genetic information is encoded in double-stranded DNA in the form of Watson-Crick base-pairing—in which adenine (A) pairs with thymine (T) and cytosine (C) with guanine (G). Depending on which set of this genetic information is decoded through transcription and translation, the developmental program and physiological status will be determined. Development of molecules that can be tailor-designed to bind sequence-specifically to any part of this genetic biopolymer (DNA or RNA), thereby enabling the control of the flow of genetic information and assessment and manipulation of the genome's structures and functions, is important for biological and biomedical research in the effort to unravel the molecular basis of life, including molecular tools for basic research in biology. This effort is also important for medicinal and therapeutic applications for the treatment and detection of genetic diseases.

[0005] Compared to proteins, RNA molecules are easier to target because they are made up of just four building blocks (A, C, G, U), whose interactions are defined by the well-established rules of Watson-Crick base-pairing. Compared to standard, double-stranded DNA (or RNA), the secondary structures of RNA are generally thermodynamically less stable and, thus, energetically less demanding for binding because, in addition to being canonical (perfectly-

matched) base-pairs, many of them are noncanonical (mismatched) and contain single-stranded loops, bulges, and junctions. The presence of these local interacting domains is essential for 'tertiary' interactions and assembly of the secondary structures into compact three-dimensional shapes. As such, slight variations in the interaction patterns or bonding strengths within these regions will have a profound effect on the overall three-dimensional folding patterns of RNA. Thus, molecules that can be used to modulate RNA interactions and thereby interfere with the RNA folding behaviors are important as molecular tools for assessing RNA functions, as well as therapeutic and diagnostic reagents.

[0006] RNA-RNA and RNA-protein interactions play key roles in gene regulation, including replication, translation, folding and packaging. The ability to selectively bind to regions within the secondary structures of RNA will often modify their physiological functions.

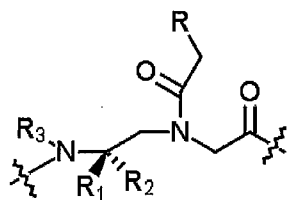
SUMMARY

[0007] Provided herein are reagents that can be used to target double-stranded nucleic acid sequences, including RNA secondary structures and mismatches. The reagents are relatively small in size, can be manufactured in large quantity and more cheaply using solution-phase methodology, and are readily taken-up by cells. They are especially appealing for targeting rapidly evolving sites, such as those associated with the pathology of cancer, bacterial and viral infection, because the described recognition scheme is modular in nature and can be readily modified to match the newly emerged sequence at will. This is a niche that is not currently fulfilled by small-molecule drugs, or traditional antisense or antigene targeting approach.

[0008] "Janus" nucleobases (JBs) are described herein. Janus nucleobases are capable of forming directional hydrogen bonding interactions with both strands of the DNA and/or RNA double helix whether or not mismatches are present. This platform has applications in basic research in biology and biotechnology, diagnostics and therapeutics. In one embodiment, the nucleobases are attached to a γ PNA backbone, integrating conformational preorganization inherent in the backbone of γ PNA with improvements in hydrogen-bonding and base-stacking capabilities, a common property of the Janus nucleobases described herein, into a single system. Unlike the natural nucleobases, which can only hybridize to one strand of the DNA or RNA double helix, these 'Janus' nucleobases can hybridize to both strands of the DNA or RNA targets. This method and platform can affect the regulation of gene expression and modulation of nucleic acid interactions—as molecular tools for basic research as well as

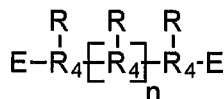
therapeutic and diagnostic reagents for the treatment and detection of genetic diseases and pathogenic infections.

[0009] Provided herein is a genetic recognition reagent, such as a nucleic acid, gamma peptide nucleic acid (γ PNA) or other nucleic acid. The genetic recognition reagent comprises a plurality of nucleobase residues attached to a nucleic acid or nucleic acid analog backbone. At least one nucleobase is a divalent nucleobase chosen from the nucleobases described in Table A, below, in which each instance of R1 is, independently, a protecting group or H and X is CH or N. Non-limiting examples of protecting groups include: methyl, formyl, ethyl, acetyl, anisyl, benzyl, benzoyl, carbamate, trifluoroacetyl, diphenylmethyl, triphenylmethyl, N-hydroxysuccinimide, benzyloxymethyl, benzyloxycarbonyl, 2-nitrobenzoyl, *t*-Boc (*tert*-butyloxycarbonyl), 4-methylbenzyl, 4-nitrophenyl, 2-chlorobenzyloxycarbonyl, 2-bromobenzyloxycarbonyl, 2,4,5-trichlorophenyl, thioanisyl, thiocresyl, cbz (carbobenzyloxy), *p*-methoxybenzyl carbonyl, 9-fluorenylmethyloxycarbonyl, pentafluorophenyl, *p*-methoxybenzyl, 3,4-dimethoxybenzyl, *p*-methoxyphenyl, 4-toluenesulfonyl, *p*-nitrobenzenesulfonates, 9-fluorenylmethyloxycarbonyl, 2-nitrophenylsulfenyl, 2,2,5,7,8-pentamethyl-chroman-6-sulfonyl, and *p*-bromobenzene sulfonyl. In certain embodiments, the backbone is chosen from one of a DNA, RNA, peptide nucleic acid (PNA), phosphorothioate, locked nucleic acid, unlocked nucleic acid, 2'-O-methyl-substituted RNA, morpholino nucleic acid, threose nucleic acid, or glycol nucleic acid backbone, or any combination thereof. In one embodiment, the backbone is a peptide nucleic acid (PNA) backbone, for example and without limitation a γ PNA backbone. The backbone is optionally PEGylated, with one or more PEG moieties of two to fifty (-O-CH₂-CH₂-) residues. An exemplary γ PNA backbone includes a backbone monomer residue that is



, where R1, R2 and R3 are, independently, H, amino acid side chains, linear or branched (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₁-C₈)hydroxyalkyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene, (C₃-C₈)cycloalkyl(C₁-C₆)alkylene, PEGylated moieties of the preceding comprising from 1 to 50 (-O-CH₂-CH₂-) residues, -CH₂-(OCH₂-CH₂)_qOP₁, -CH₂-(OCH₂-CH₂)_q-NHP₁, -CH₂-(OCH₂-CH₂-O)_q-SP₁ and -CH₂-(SCH₂-CH₂)_q-SP₁, -CH₂-(OCH₂-CH₂)_r-OH, -CH₂-(OCH₂-CH₂)_r-NH₂, -CH₂-(OCH₂-CH₂)_r-NHC(NH)NH₂, or -CH₂-(OCH₂-CH₂)_r-S-S-[CH₂CH₂]_s-NHC(NH)NH₂, where P₁ is

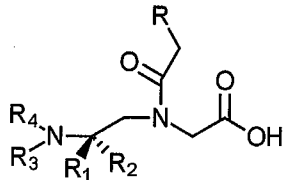
selected from the group consisting of H, (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene and (C₃-C₈)cycloalkyl(C₁-C₆)alkylene; q is an integer from 0 to 10, inclusive; r and s are each independently integers from 1 to 50, inclusive; where R₁ and R₂ are different and one of R₁ or R₂ is H. In one example R₂ is H, R₁ is an amino acid side chain that is optionally PEGylated, with one or more PEG moieties of one to twelve (-O-CH₂-CH₂-) residues. The nucleobases are arranged in certain embodiments in a sequence complementary to a target sequence of a nucleic acid. In one embodiment, the genetic recognition reagent has 3 to 25 nucleobases. In one embodiment of the composition described above the genetic recognition reagent has the



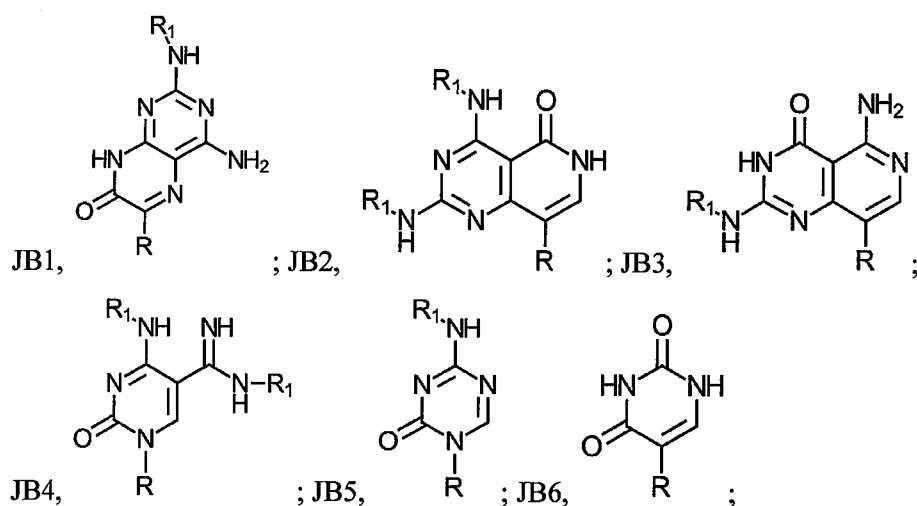
structure: , where each instance of R₄ is a backbone monomer residue and each instance of R is a nucleobase, where at least one instance of R the divalent nucleobase, E are independently end groups, and "n" is zero or a positive integer ranging from 1 to 48, in which the sequence of nucleobases R is complementary to a target sequence of a nucleic acid. In another embodiment the divalent nucleobases are, independently chosen from JB₁, JB₂, JB₃ and JB₄. In yet another embodiment, all instances of R₁ are H.

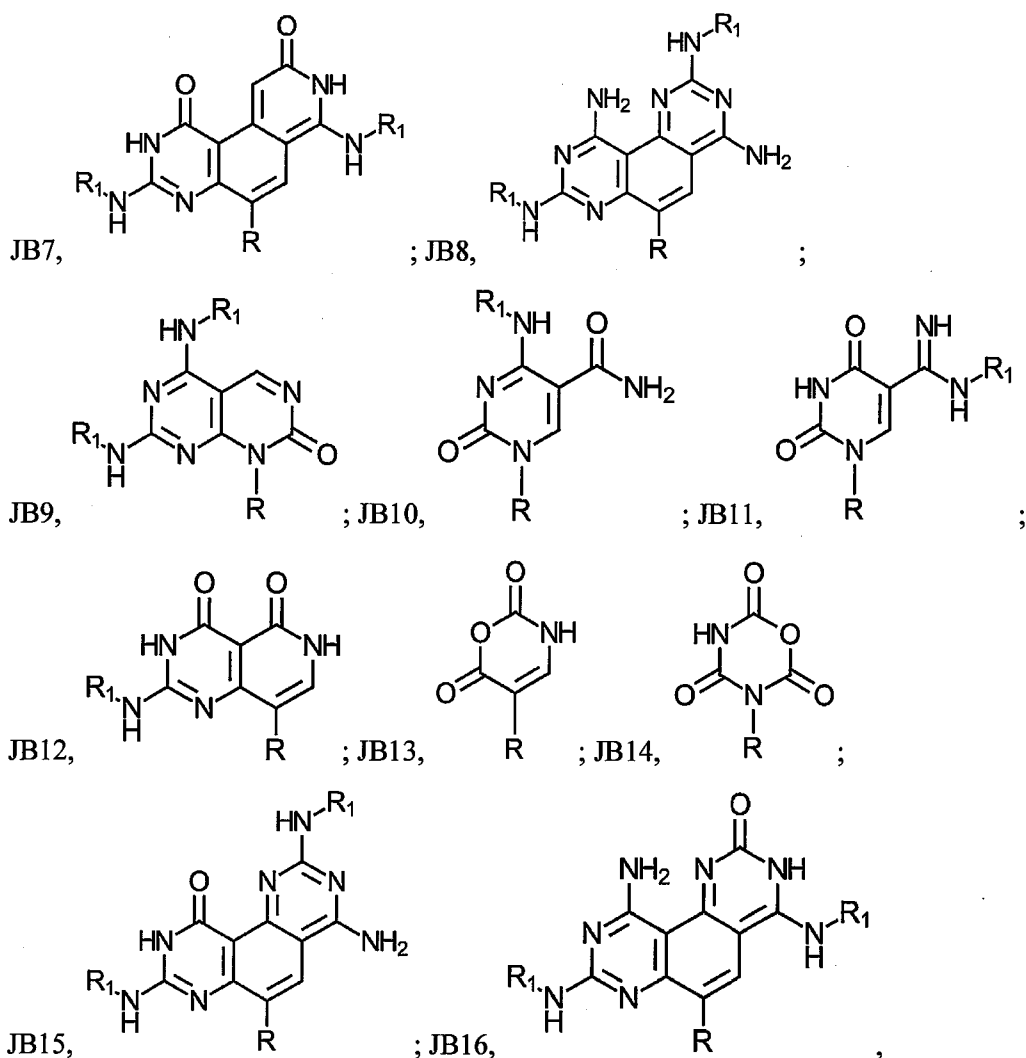
[0010] Also provided is a monomer for production of a genetic recognition reagent comprising a backbone monomer for a genetic recognition reagent covalently attached to a divalent nucleobase chosen from those listed in Table A, in which each instance of R₁ is, independently, a protecting group or H and X is CH or N. In one example, the nucleobase is one of JB₁, JB₂, JB₃ or JB₄. In another embodiment, R₁ is a protecting group, such as, for example and without limitation, a protecting group is chosen from one or more of: methyl, formyl, ethyl, acetyl, anisyl, benzyl, benzoyl, carbamate, trifluoroacetyl, diphenylmethyl, triphenylmethyl, N-hydroxysuccinimide, benzyloxymethyl, benzyloxycarbonyl, 2-nitrobenzoyl, *t*-Boc (*tert*-butyloxycarbonyl), 4-methylbenzyl, 4-nitrophenyl, 2-chlorobenzyloxycarbonyl, 2-bromobenzyloxycarbonyl, 2,4,5-trichlorophenyl, thioanisyl, thiocresyl, cbz (carbobenzyloxy), *p*-methoxybenzyl carbonyl, 9-fluorenylmethyloxycarbonyl, pentafluorophenyl, *p*-methoxybenzyl, 3,4-dimethoxybenzyl, *p*-methoxyphenyl, 4-toluenesulfonyl, *p*-nitrobenzenesulfonates, 9-fluorenylmethyloxycarbonyl, 2-nitrophenylsulfonyl, 2,2,5,7,8-pentamethyl-chroman-6-sulfonyl, and *p*-bromobenzenesulfonyl. In one embodiment, the monomer is a peptide nucleic acid (PNA) monomer, and in another, γ PNA. In yet another embodiment, the backbone monomer is a PNA or γ PNA that is PEGylated, with one or more PEG moieties of two to fifty (-O-CH₂-

CH₂-) residues. In a further embodiment, the monomer is a γ PNA monomer having the

structure: , where R1, R2 and R4 are, independently, H, amino acid side chains, linear or branched (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₁-C₈)hydroxyalkyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene, (C₃-C₈)cycloalkyl(C₁-C₆)alkylene, PEGylated moieties of the preceding comprising from 1 to 50 (-O-CH₂-CH₂-) residues, -CH₂-(OCH₂-CH₂)_qOP₁, -CH₂-(OCH₂-CH₂)_q-NHP₁, -CH₂-(OCH₂-CH₂-O)_q-SP₁ and -CH₂-(SCH₂-CH₂)_q-SP₁, -CH₂-(OCH₂-CH₂)_r-OH, -CH₂-(OCH₂-CH₂)_r-NH₂, -CH₂-(OCH₂-CH₂)_r-NHC(NH)NH₂, or -CH₂-(OCH₂-CH₂)_r-S-S[CH₂CH₂]_sNHC(NH)NH₂, where P₁ is selected from the group consisting of H, (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene and (C₃-C₈)cycloalkyl(C₁-C₆)alkylene; q is an integer from 0 to 10, inclusive; r and s are each independently integers from 1 to 50, inclusive; where R1 and R2 are different, and one of R1 or R2 is H, and R3 is H or a protecting group. In one non-limiting example, R2 is H, R1 is an amino acid side chain that is optionally PEGylated, with one or more PEG moieties of one to twelve (-O-CH₂-CH₂-) residues, and R3 is a protecting group.

[0011] In another embodiment, a divalent nucleobase is provided, chosen from the divalent nucleobases of Table A, in which each instance of R₁ is, independently a protecting group or H and R is a reactive group. In another embodiment, the nucleobase is chosen from one the following:





in which each instance of R1 is, independently a protecting group or H and R is carboxyl.

[0012] In another embodiment, a kit is provided comprising any monomer as described herein in a vessel. In one embodiment, the kit comprises monomers comprising each of JB1, JB2, JB3, JB4, JB5, JB6, JB7, JB8, JB9, JB9b, JB10, JB11, JB12, JB13, JB14, JB15, JB16 nucleobases, with each different monomer in a separate vessel. An array comprising a genetic recognition reagent as described herein also is provided.

[0013] A method of detection of a target sequence in a nucleic acid sample is provided. The method comprising contacting a genetic recognition reagent described herein with a sample comprising nucleic acid and detecting binding of the genetic recognition reagent with a nucleic acid.

[0014] Lastly, a method of isolation and purification of a nucleic acid containing a target sequence is provided. The method comprising, contacting a nucleic acid sample with a

genetic recognition reagent as described herein, separating the nucleic acid sample from the genetic recognition reagent, leaving any nucleic acid bound to the genetic recognition reagent bound to the genetic recognition reagent, and separating the genetic recognition reagent from any nucleic acid bound to the genetic recognition reagent. In one embodiment of the method, the genetic recognition reagent is immobilized on a substrate. The method comprising contacting a nucleic acid with the substrate, washing the substrate to remove unbound nucleic acid from the substrate, but leaving bound nucleic acid bound to the substrate, and eluting the bound nucleic acid from the substrate.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Figure 1 illustrates hydrogen-bonding interactions between (A) natural base-pairs, (B) JB 1-4 (labeled 1-4) and the perfectly-matched DNA or RNA target and (C) JB 5-16 (labeled 5-16) and the mismatched DNA or RNA target.

[0016] Figure 2 provides structures of exemplary nucleobases.

[0017] Figures 3A-3F provide exemplary structures for nucleic acid analogs.

[0018] Figure 4 provides examples of amino acid side chains.

[0019] Figure 5 illustrates simulated structures γ PNAJB1-4 invading DNA double helix.

[0020] Figure 6 illustrates simulated structures of γ PNAJB1-16 invading an RNA secondary structure containing perfectly-matched and mismatched base-pairs.

[0021] Figure 7 illustrates exemplary schemes for synthesis of the described divalent nucleobases.

[0022] Figure 8 illustrates an exemplary scheme for the synthesis of γ PNA monomers with divalent nucleobases.

[0023] Figure 9 illustrates an exemplary scheme for synthesis of γ PNA oligomers with divalent nucleobases.

DETAILED DESCRIPTION

[0024] The use of numerical values in the various ranges specified in this application, unless expressly indicated otherwise, are stated as approximations as though the minimum and maximum values within the stated ranges are both preceded by the word "about". In this manner, slight variations above and below the stated ranges can be used to achieve substantially the same results as values within the ranges. Also, unless indicated otherwise, the disclosure of ranges is intended as a continuous range including every value between the minimum and maximum values. As used herein "a" and "an" refer to one or more.

[0025] As used herein, the term “comprising” is open-ended and may be synonymous with “including”, “containing”, or “characterized by”. The term “consisting essentially of” limits the scope of a claim to the specified materials or steps and those that do not materially affect the basic and novel characteristic(s) of the claimed invention. The term “consisting of” excludes any element, step, or ingredient not specified in the claim. As used herein, embodiments “comprising” one or more stated elements or steps also include, but are not limited to embodiments “consisting essentially of” and “consisting of” these stated elements or steps.

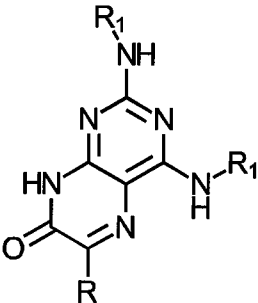
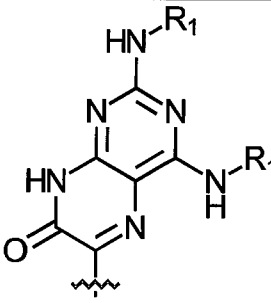
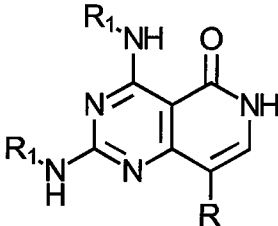
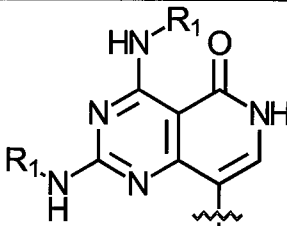
[0026] Provided herein are nucleic acids and analogs thereof, collectively “genetic recognition reagents” (genetic recognition reagent), that bind specifically to two nucleic acid strands, whether or not the two strands are independent strands, two portions of a single strand (e.g., in a hairpin) or contain mismatches in the sense that at one or more positions within the two strands at the site of binding to the genetic recognition reagents, the bases are not able to base pair according to traditional Watson-Crick base pairing (A-T/U, T/U-A, G-C or C-G). In one embodiment, the two strands binding the genetic recognition reagent are non-complementary, meaning they do not hybridize under physiological conditions and typically contain less than 50% complementarity, meaning that less than 50% of the bases in the two strands are mismatched when aligned to nucleobases of the genetic recognition reagent. Thus, depending upon choice of nucleobases in the sequence, the genetic recognition reagents described herein can invade or otherwise hybridize to two strands of fully-complementary, partially-complementary or non-complementary double-stranded nucleic acids.

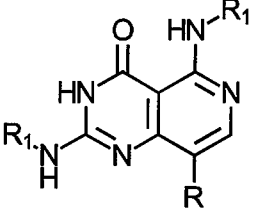
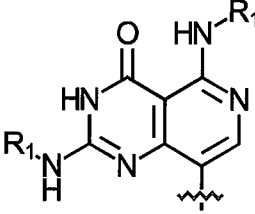
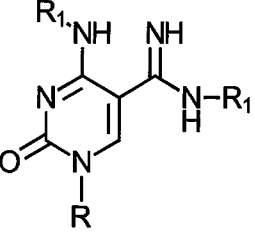
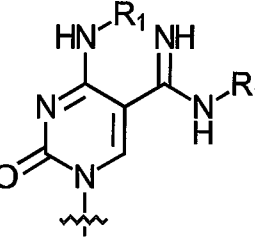
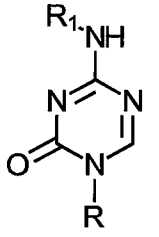
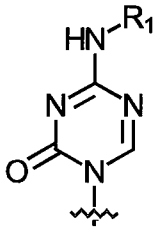
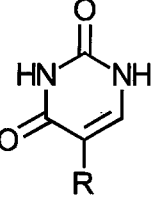
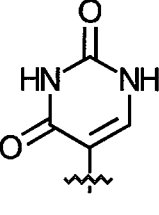
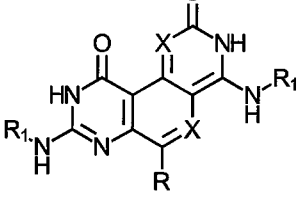
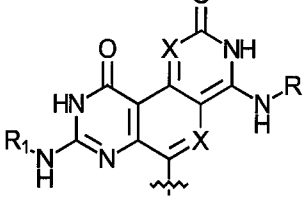
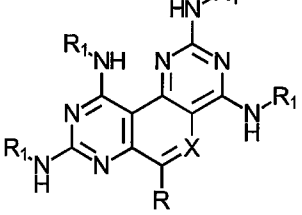
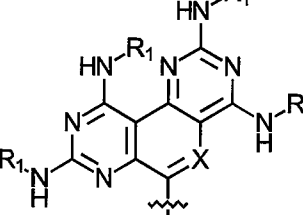
[0027] In one embodiment, the genetic recognition reagents described herein comprise all divalent nucleobases. In another embodiment, the genetic recognition reagents described herein comprise at least one divalent nucleobases, with other nucleobases being monovalent. As used herein, a monovalent nucleobase binds one nucleobase on a single nucleic acid strand, while a divalent nucleobase, e.g., a Janus base described herein, binds to two nucleobases, one on a first nucleic acid strand, and another on a second nucleic acid strand.

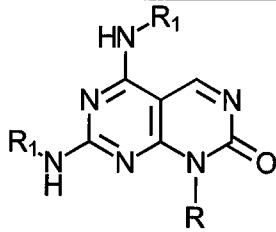
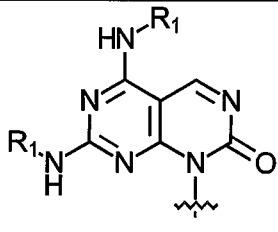
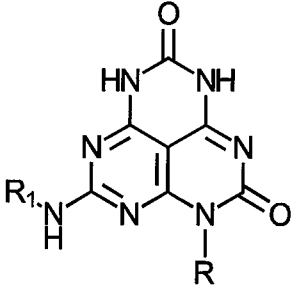
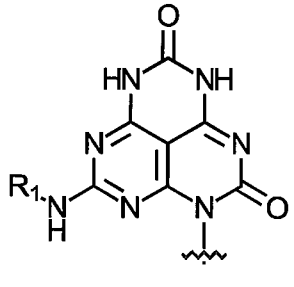
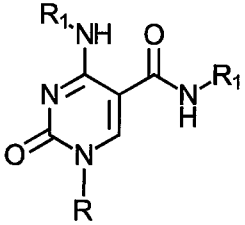
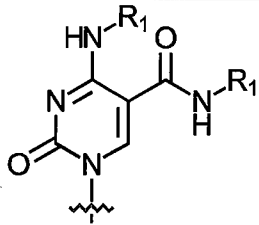
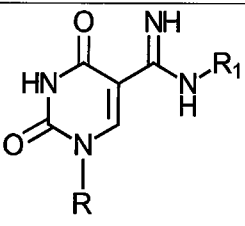
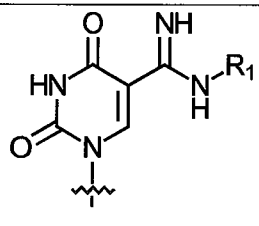
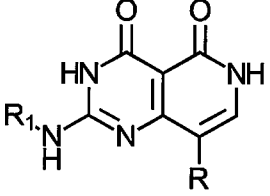
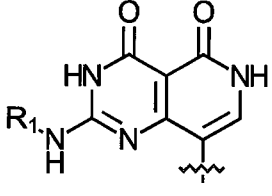
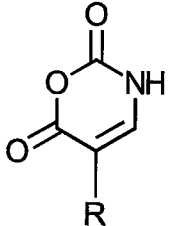
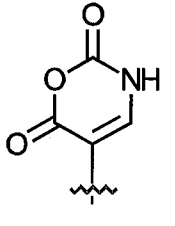
[0028] Thus in one embodiment, divalent nucleobases are provided. Those nucleobases can be incorporated into a genetic recognition reagent monomer, which can then be incorporated into an oligomer of monomers with a desired sequence of nucleobases. Table A provides exemplary divalent “Janus” bases, their binding specificities (*see also*,

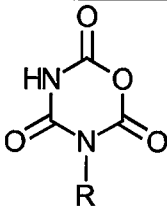
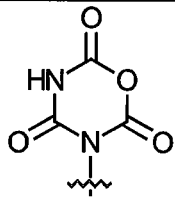
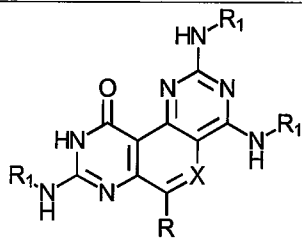
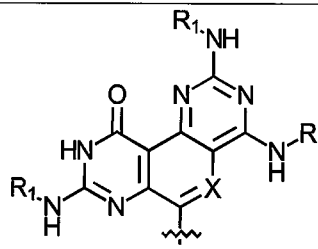
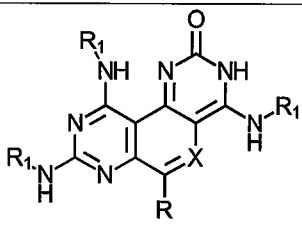
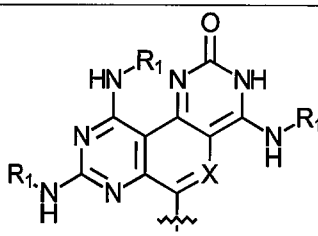
Figure 1), structures for the nucleobases, and structures for the nucleobases residue as it is attached to a nucleotide and/or a genetic recognition reagent. In the context of the present disclosure, a “nucleotide” refers to a monomer comprising at least nucleobases and a backbone element, which in a nucleic acid, such as RNA or DNA is ribose or deoxyribose. “Nucleotides” also typically comprise reactive groups that permit polymerization under specific conditions. In native DNA and RNA, those reactive groups are the 5’ phosphate and 3’ hydroxyl groups. For chemical synthesis of nucleic acids and analogs thereof, the bases and backbone monomers may contain modified groups, such as blocked amines, as are known in the art. A “nucleotide residue” refers to a single nucleotide that is incorporated into an oligonucleotide or polynucleotide. Likewise, a “nucleobases residue” refers to a nucleobases incorporated into a nucleotide or a nucleic acid or analog thereof. A “genetic recognition reagent” refers generically to a nucleic acid or a nucleic acid analog that comprises a sequence of nucleobases that is able to hybridize to a complementary nucleic acid sequence on a nucleic acid by cooperative base pairing, *e.g.*, Watson-Crick base pairing or Watson-Crick-like base pairing. Of note, JB1-JB4 bind complementary bases (C-G, G-C, A-T and T-A), while JB5-JB16 bind mismatches, and thus can be used to bind two strands of matched and/or mismatched bases.

Table A - Divalent Nucleobases

Nucleobase	Bases represented	Nucleobase	Nucleobase residue
JB1	T/D*		
JB2	D/T		

JB3	G/C		
JB4	C/G		
JB5	C/C		
JB6	U/U		
JB7	G/G		
JB8	D/D		

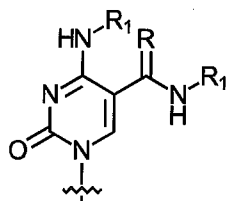
JB9	A/C		
JB9b	A/C		
JB10	C/A		
JB11	U/G		
JB12	G/U		
JB13	C/U		

JB14	U/C		
JB15	G/D		
JB16	D/G		

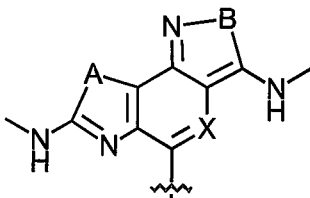
*diaminopurine, an adenine analog

[0029] In Table A, R is a reactive group that reacts with a backbone monomer during synthesis of a monomer. Non-limiting examples of a reactive group is a carboxyl (e.g., $-\text{C}(\text{O})\text{OH}$), hydroxyl (e.g., $-\text{C}-\text{OH}$), cyanate (e.g., $-\text{C}-\text{C}\equiv\text{N}$), thiol (e.g., $-\text{C}-\text{SH}$), $(\text{CH}_2)_n\text{CO}_2\text{H}$ or $(\text{CH}_2)_n\text{CO}_2\text{Y}$ ($n=1-5$, Y= any leaving group such as Cl, alkyl, aryl, *etc.*). X is either CH or N. R1s are each, independently: H or a protecting group. In one embodiment, all instances of X are C ($=\text{CH}-$) and all instances are R1. Where all instances of R1 are H, the compounds or moieties are said to be deprotected. Depending on the chemistries employed to prepare the monomers or oligomers comprising the monomers, one or more of R1 is optionally protected using a protecting group, as is needed. Protecting groups for amines, include, for example and without limitation: methyl, formyl, ethyl, acetyl, anisyl, benzyl, benzoyl, carbamate, trifluoroacetyl, diphenylmethyl, triphenylmethyl, N-hydroxysuccinimide, benzyloxymethyl, benzyloxycarbonyl, 2-nitrobenzoyl, *t*-Boc (*tert*-butyloxycarbonyl), 4-methylbenzyl, 4-nitrophenyl, 2-chlorobenzyloxycarbonyl, 2-bromobenzyloxycarbonyl, 2,4,5-trichlorophenyl, thioanisyl, thiocresyl, cbz (carbobenzyloxy), *p*-methoxybenzyl carbonyl, 9-fluorenylmethyloxycarbonyl, pentafluorophenyl, *p*-methoxybenzyl, 3,4-dimethoxybenzyl, *p*-methoxyphenyl, 4-toluenesulfonyl, *p*-nitrobenzenesulfonates, 9-fluorenylmethyloxycarbonyl,

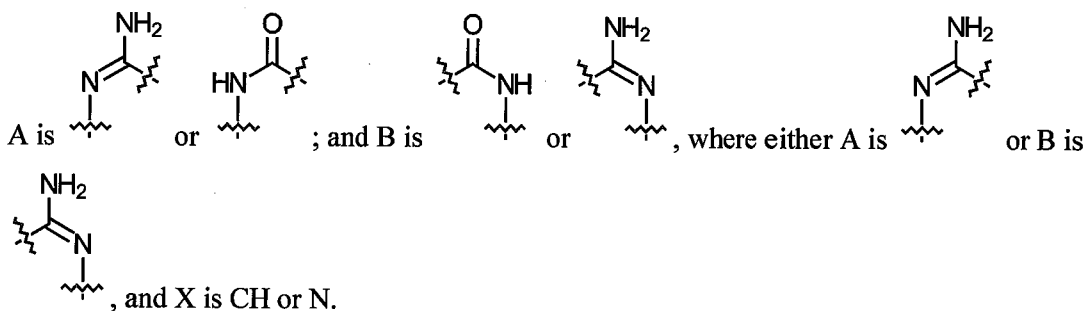
2-nitrophenylsulfonyl, 2,2,5,7,8-pentamethyl-chroman-6-sulfonyl, and *p*-bromobenzenesulfonyl. JB9b is a structural variation of JB9.



[0030] The structure, is generic to JB4 and JB10, wherein R is



NH or O . The structure, , is generic to JB8, JB15 and JB 16, where:



[0031] The nucleobases of Table A have divalent binding affinity, as indicated. Figure 1 depicts the hydrogen bonding of the sixteen divalent nucleobases of Table A, while comparing to Watson-Crick-like hydrogen-bonding interactions with natural bases. Of note, JB1, JB2, JB3 and JB4 account for matched (complementary) sequences, while the remainder of the compounds bind all other iterations of mis-matched sequences.

[0032] Nucleobases are recognition moieties that bind specifically to one or more of adenine, guanine, thymine, cytosine, and uracil, *e.g.*, by Watson-Crick or Watson-Crick-like base pairing by hydrogen bonding. A “nucleobase” includes primary nucleobases: adenine, guanine, thymine, cytosine, and uracil, as well as modified purine and pyrimidine bases, such as, without limitation, hypoxanthine, xanthene, 7-methylguanine, 5, 6, dihydrouracil, 5-methylcytosine, and 5-hydroxymethylcytosine. Figure 2 also depicts non-limiting examples of nucleobases, including monovalent nucleobases (*e.g.*, adenine, cytosine, guanine, thymine or uracil, which bind to one strand of nucleic acid or nucleic acid analogs), and “clamp” nucleobases, such as a “G-clamp,” which binds complementary nucleobases with enhanced strength. Additional purine, purine-like, pyrimidine and pyrimidine-like nucleobases are

known in the art, for example as disclosed in United States Patent Nos. 8,053,212, 8,389,703, and 8,653,254.

[0033] Also provided herein are nucleotides having the structure A-B wherein A is a backbone monomer and B is a divalent nucleobase as described above. The backbone monomer can be any suitable nucleic acid backbone monomer, such as a ribose triphosphate or deoxyribose triphosphate, or a monomer of a nucleic acid analog, such as peptide nucleic acid (PNA), such as a gamma PNA (γ PNA). The backbone monomer includes both the structural “residue” component, such as the ribose in RNA, and any active groups that are modified in linking monomers together, such as the 5' triphosphate and 3' hydroxyl groups of a ribonucleotide, which are modified when polymerized into RNA to leave a phosphodiester linkage. Likewise for PNA, the C-terminal carboxyl and N-terminal amine active groups of the N-(2-aminoethyl)glycine backbone monomer are condensed during polymerization to leave a peptide (amide) bond. In another embodiment, the active groups are phosphoramidite groups useful for phosphoramidite oligomer synthesis, as is broadly-known in the arts. The nucleotide also optionally comprises one or more protecting groups as are known in the art, such as 4,4'-dimethoxytrityl (DMT), and as described herein. A number of additional methods of preparing synthetic genetic recognition reagents are known, and depend on the backbone structure and particular chemistry of the base addition process. Determination of which active groups to utilize in joining nucleotide monomers and which groups to protect in the bases, and the required steps in preparation of oligomers is well within the abilities of those of ordinary skill in the chemical arts and in the particular field of nucleic acid and nucleic acid analog oligomer synthesis.

[0034] Non-limiting examples of common nucleic acid analogs include peptide nucleic acids, such as γ PNA, phosphorothioate (e.g., Figure 3A), locked nucleic acid (2'-O-4'-C-methylene bridge, including oxy, thio or amino versions thereof, e.g., Figure 3B), unlocked nucleic acid (the C2'-C3' bond is cleaved, e.g., Figure 3C), 2'-O-methyl-substituted RNA, morpholino nucleic acid (e.g., Fig _D), threose nucleic acid (e.g., Figure 3E), glycol nucleic acid (e.g., Figure 3F, showing *R* and *S* Forms), *etc.* Figure 3A-F shows monomer structures for various examples of nucleic acid analogs. Figures 3A-3F each show two monomer residues incorporated into a longer chain as indicated by the wavy lines. Incorporated monomers are referred to herein as “residues” and the part of the nucleic acid or nucleic acid analog oligomer or polymer excluding the nucleobases is referred to as the “backbone” of the nucleic acid or nucleic acid analog. As an example, for RNA, an exemplary nucleobase is adenine, a corresponding monomer is adenosine triphosphate, and the incorporated residue is an

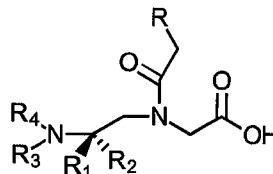
adenosine monophosphate residue. For RNA, the “backbone” consists of ribose subunits linked by phosphates, and thus the backbone monomer is ribose triphosphate prior to incorporation and a ribose monophosphate residue after incorporation.

[0035] According to one embodiment, with the advent of conformationally-preorganized γ PNA, precise sequence selection is no longer an issue (Bahal, R., *et al.* “Sequence-unrestricted, Watson-Crick recognition of double helical B-DNA by (R)-MiniPEG- γ PNAs (2012) *ChemBioChem* 13:56-60). γ PNA can be designed to bind to any sequence of double helical B-DNA based on the well-established rules of Watson-Crick base-pairing. However, with an arsenal of only natural nucleobases as recognition elements, strand invasion of DNA by γ PNA is confined to sub-physiological ionic strengths (Rapireddy, S., R. *et al.* “Strand invasion of mixed-sequence, double-helical B-DNA by γ -peptide nucleic acids containing G-clamp nucleobases under physiological conditions,” (2011) *Biochemistry* 50:3913-3918). A reduction in the efficiency of productive γ PNA binding at elevated ionic strengths is not due to the lack of base-pair accessibility, but rather due to the lack of binding free energy. Under physiological conditions, DNA double helix is sufficiently dynamic to permit strand invasion, provided that the required binding free energy could be met. One way to improve the binding free energy of such a system would be to enhance the base-stacking and H-bonding capabilities of the recognition elements, which is met by the present invention. In one embodiment, γ PNA monomers and oligomers containing a specialized set of divalent “Janus” nucleobases (JBs) that are capable of forming directional hydrogen bonding interactions with both strands of the DNA or RNA double helix is performed. Examples of the chemical structures of the JBs are illustrated in Table A. Non-limiting examples of γ PNA monomers and oligomers are provided below, with, e.g., an amino acid side chain, or a PEGylated (polyethyleneglycol, or PEG) group at the chiral gamma carbon.

[0036] As used herein, the term “nucleic acid” refers to deoxyribonucleic acids (DNA) and ribonucleic acids (RNA). Nucleic acid analogs include, for example and without limitation: 2'-O-methyl-substituted RNA, locked nucleic acids, unlocked nucleic acids, triazole-linked DNA, peptide nucleic acids, morpholino oligomers, dideoxynucleotide oligomers, glycol nucleic acids, threose nucleic acids and combinations thereof including, optionally ribonucleotide or deoxyribonucleotide residue(s). Herein, “nucleic acid” and “oligonucleotide”, which is a short, single-stranded structure made of up nucleotides, are used interchangeably. An oligonucleotide may be referred to by the length (i.e. number of

nucleotides) of the strand, through the nomenclature “-mer”. For example, an oligonucleotide of 22 nucleotides would be referred to as a 22-mer.

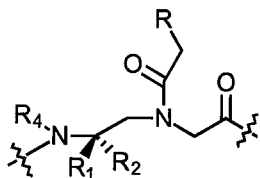
[0037] A "peptide nucleic acid" refers to a DNA or RNA analog or mimic in which the sugar phosphodiester backbone of the DNA or RNA is replaced by a N-(2-aminoethyl)glycine unit. A gamma PNA (γ PNA) is an oligomer or polymer of gamma-modified N-(2-



aminoethyl)glycine monomers of the following structure: , where at

least one of R1 or R2 attached to the gamma carbon is not a hydrogen, or R1 and R2 are different, such that the gamma carbon is a chiral center. When R1 and R2 are hydrogen (N-(2-aminoethyl)-glycine backbone), or the same, there is no such chirality about the gamma carbon. When R1 and R2 are different, such as when one of R1 or R2 are H and the other is not, there is chirality about the gamma carbon. Typically, for γ PNAs and γ PNA monomers, either of R1 or R2 is an H and the other is an amino acid sidechain or an organic group, such as a (C₁-C₁₀) organic group or hydrocarbon, optionally PEGylated with from 1 to 50 oxyethylene residues - that is, [-O-CH₂-CH₂-]_n, where n is 1 to 50, inclusive. R4 can be H or an organic group, such as a (C₁-C₁₀) organic group or hydrocarbon, optionally PEGylated with from 1 to 50 oxyethylene residues. For example and without limitation, R1, R2 and R4 are, independently, H, amino acid side chains, linear or branched (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₁-C₈)hydroxyalkyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene, (C₃-C₈)cycloalkyl(C₁-C₆)alkylene, PEGylated moieties of the preceding comprising from 1 to 50 (-O-CH₂-CH₂-) residues, -CH₂-(OCH₂-CH₂)_qOP₁, -CH₂-(OCH₂-CH₂)_q-NHP₁, -CH₂-(OCH₂-CH₂-O)_q-SP₁ and -CH₂-(SCH₂-CH₂)_q-SP₁, -CH₂-(OCH₂-CH₂)_r-OH, -CH₂-(OCH₂-CH₂)_r-NH₂, -CH₂-(OCH₂-CH₂)_r-NHC(NH)NH₂, or -CH₂-(OCH₂-CH₂)_r-S-S[CH₂CH₂]_sNHC(NH)NH₂, where P₁ is selected from the group consisting of H, (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene and (C₃-C₈)cycloalkyl(C₁-C₆)alkylene; q is an integer from 0 to 10, inclusive; r and s are each independently integers from 1 to 50, inclusive; where R1 and R2 are different, and optionally one of R1 or R2 is H. R3 is H or a protective group

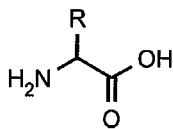
[0038] A γ PNA monomer incorporated into a γ PNA oligomer or polymer,



is referred to herein as a “ γ PNA monomer residue”, with each residue

having the same or different R group as its nucleobase, such as adenine, guanine, cytosine, thymine and uracil bases, or other nucleobases, such as the monovalent and divalent bases described herein, such that the order of bases on the γ PNA is its “sequence”, as with DNA or RNA. The depicted γ PNA monomer and residue structures show a backbone monomer, and a backbone monomer residue, respectively, attached to a nucleobase (R). A sequence of nucleobases in a nucleic acid or a nucleic acid analog oligomer or polymer, such as a γ PNA oligomer or polymer, binds to a complementary sequence of adenine, guanine, cytosine, thymine and/or uracil residues in a nucleic acid strand by cooperative bonding, essentially as with Watson-Crick binding of complementary bases in double-stranded DNA or RNA. “Watson-Crick-like” bonding refers to hydrogen bonding of nucleobases other than G, A, T, C or U, such as the bonding of the divalent bases shown herein with G, A, T, C, U or other nucleobases.

[0039] An “amino acid side chain” is a side chain for an amino acid. Amino acids have the



structure: $\text{H}_2\text{N}-\text{CH}(\text{R})-\text{COOH}$, where R is the amino acid side chain. Non-limiting examples of amino acid side chains are shown in Figure 4. Glycine is not represented because in the embodiment R1s are both H.

[0040] The following are exemplary definitions of various moieties or groups as used herein.

“Alkyl” refers to straight, branched chain, or cyclic hydrocarbon groups including from 1 to about 20 carbon atoms, for example and without limitation C_{1-3} , C_{1-6} , C_{1-10} groups, for example and without limitation, straight, branched chain alkyl groups such as methyl, ethyl, propyl, butyl, pentyl, hexyl, heptyl, octyl, nonyl, decyl, undecyl, dodecyl, and the like.

“Substituted alkyl” refers to alkyl substituted at 1 or more, e.g., 1, 2, 3, 4, 5, or even 6 positions, which substituents are attached at any available atom to produce a stable compound, with substitution as described herein. “Optionally substituted alkyl” refers to alkyl or substituted alkyl. “Halogen,” “halide,” and “halo” refers to -F, -Cl, -Br, and/or -I.

“Alkylene” and “substituted alkylene” refer to divalent alkyl and divalent substituted alkyl,

respectively, including, without limitation, ethylene (-CH₂-CH₂-). "Optionally substituted alkylene" refers to alkylene or substituted alkylene.

[0041] "Alkene or alkenyl" refers to straight, branched chain, or cyclic hydrocarbyl groups including from 2 to about 20 carbon atoms, such as, without limitation C₁₋₃, C₁₋₆, C₁₋₁₀ groups having one or more, e.g., 1, 2, 3, 4, or 5, carbon-to-carbon double bonds. "Substituted alkene" refers to alkene substituted at 1 or more, e.g., 1, 2, 3, 4, or 5 positions, which substituents are attached at any available atom to produce a stable compound, with substitution as described herein. "Optionally substituted alkene" refers to alkene or substituted alkene. Likewise, "alkenylene" refers to divalent alkene. Examples of alkenylene include without limitation, ethenylene (-CH=CH-) and all stereoisomeric and conformational isomeric forms thereof. "Substituted alkenylene" refers to divalent substituted alkene. "Optionally substituted alkenylene" refers to alkenylene or substituted alkenylene.

[0042] "Alkyne or alkynyl" refers to a straight or branched chain unsaturated hydrocarbon having the indicated number of carbon atoms and at least one triple bond. Examples of a (C₂-C₈)alkynyl group include, but are not limited to, acetylene, propyne, 1-butyne, 2-butyne, 1-pentyne, 2-pentyne, 1-hexyne, 2-hexyne, 3-hexyne, 1-heptyne, 2-heptyne, 3-heptyne, 1-octyne, 2-octyne, 3-octyne and 4-octyne. An alkynyl group can be unsubstituted or optionally substituted with one or more substituents as described herein below. The term "alkynylene" refers to divalent alkyne. Examples of alkynylene include without limitation, ethynylene, propynylene. "Substituted alkynylene" refers to divalent substituted alkyne.

[0043] The term "alkoxy" refers to an -O-alkyl group having the indicated number of carbon atoms. For example, a (C₁-C₆)alkoxy group includes -O-methyl (methoxy), -O-ethyl (ethoxy), -O-propyl (propoxy), -O-isopropyl (isopropoxy), -O-butyl (butoxy), -O-sec-butyl (sec-butoxy), -O-tert-butyl (tert-butoxy), -O-pentyl (pentoxy), -O-isopentyl (isopentoxy), -O-neopentyl (neopentoxy), -O-hexyl (hexyloxy), -O-isohexyl (isohexyloxy), and -O-neohexyl (neohexyloxy). "Hydroxyalkyl" refers to a (C₁-C₁₀)alkyl group wherein one or more of the alkyl group's hydrogen atoms is replaced with an -OH group. Examples of hydroxyalkyl groups include, but are not limited to, -CH₂OH, -CH₂CH₂OH, -CH₂CH₂CH₂OH, -CH₂CH₂CH₂CH₂OH, -CH₂CH₂CH₂CH₂CH₂OH, -CH₂CH₂CH₂CH₂CH₂CH₂OH, and branched versions thereof. The term "ether" or "oxygen ether" refers to (C₁-C₁₀)alkyl group wherein one or more of the alkyl group's carbon atoms is replaced with an -O- group. The term ether includes -CH₂-(OCH₂-CH₂)_qOP₁ compounds where P₁ is a protecting group, -H, or

a (C₁-C₁₀)alkyl. Exemplary ethers include polyethylene glycol, diethylether, methylhexyl ether and the like.

[0044] The term "thioether" refers to (C₁-C₁₀)alkyl group wherein one or more of the alkyl group's carbon atoms is replaced with an -S- group. The term thioether includes -CH₂-(SCH₂-CH₂)_q-SP₁ compounds where P₁ is a protecting group, -H, or a (C₁-C₁₀)alkyl. Exemplary thioethers include dimethylthioether, ethylmethyl thioether. Protecting groups are known in the art and include, without limitation: 9-fluorenylmethyloxy carbonyl (Fmoc), t-butyloxycarbonyl (Boc), benzhydryloxycarbonyl (Bhoc), benzyloxycarbonyl (Cbz), O-nitroveratryloxycarbonyl (Nvoc), benzyl (Bn), allyloxycarbonyl (alloc), trityl (Trt), dimethoxytrityl (DMT), 1-(4,4-dimethyl-2,6-dioxacyclohexylidene)ethyl (Dde), diathiasuccinoyl (Dts), benzothiazole-2-sulfonyl (Bts) and monomethoxytrityl (MMT) groups.

[0045] "Aryl," alone or in combination refers to an aromatic monocyclic or bicyclic ring system such as phenyl or naphthyl. "Aryl" also includes aromatic ring systems that are optionally fused with a cycloalkyl ring. A "substituted aryl" is an aryl that is independently substituted with one or more substituents attached at any available atom to produce a stable compound, wherein the substituents are as described herein. "Optionally substituted aryl" refers to aryl or substituted aryl. "Arylene" denotes divalent aryl, and "substituted arylene" refers to divalent substituted aryl. "Optionally substituted arylene" refers to arylene or substituted arylene.

[0046] "Heteroatom" refers to N, O, P and S. Compounds that contain N or S atoms can be optionally oxidized to the corresponding N-oxide, sulfoxide or sulfone compounds. "Hetero-substituted" refers to an organic compound in any embodiment described herein in which one or more carbon atoms are substituted with N, O, P or S.

[0047] "Cycloalkyl" refer to monocyclic, bicyclic, tricyclic, or polycyclic, 3- to 14-membered ring systems, which are either saturated, unsaturated or aromatic. The cycloalkyl group may be attached via any atom. Cycloalkyl also contemplates fused rings wherein the cycloalkyl is fused to an aryl or hetroaryl ring. Representative examples of cycloalkyl include, but are not limited to cyclopropyl, cyclobutyl, cyclopentyl, and cyclohexyl. A cycloalkyl group can be unsubstituted or optionally substituted with one or more substituents as described herein below. "Cycloalkylene" refers to divalent cycloalkyl. The term "optionally substituted cycloalkylene" refers to cycloalkylene that is substituted with 1, 2 or 3

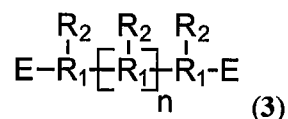
substituents, attached at any available atom to produce a stable compound, wherein the substituents are as described herein.

[0048] "Carboxyl" or "carboxylic" refers to group having the indicated number of carbon atoms and terminating in a $-C(O)OH$ group, thus having the structure $-R-C(O)OH$, where R is a divalent organic group that includes linear, branched, or cyclic hydrocarbons. Non-limiting examples of these include: C_{1-8} carboxylic groups, such as ethanoic, propanoic, 2-methylpropanoic, butanoic, 2,2-dimethylpropanoic, pentanoic, *etc.*

[0049] " (C_3-C_8) aryl- (C_1-C_6) alkylene" refers to a divalent alkylene wherein one or more hydrogen atoms in the C_1-C_6 alkylene group is replaced by a (C_3-C_8) aryl group. Examples of (C_3-C_8) aryl- (C_1-C_6) alkylene groups include without limitation 1-phenylbutylene, phenyl-2-butylene, 1-phenyl-2-methylpropylene, phenylmethylene, phenylpropylene, and naphthylethylene. The term " (C_3-C_8) cycloalkyl- (C_1-C_6) alkylene" refers to a divalent alkylene wherein one or more hydrogen atoms in the C_1-C_6 alkylene group is replaced by a (C_3-C_8) cycloalkyl group. Examples of (C_3-C_8) cycloalkyl- (C_1-C_6) alkylene groups include without limitation 1-cyclopropylbutylene, cyclopropyl-2-butylene, cyclopentyl-1-phenyl-2-methylpropylene, cyclobutylmethylene and cyclohexylpropylene.

[0050] Unless otherwise indicated, the nucleic acids and nucleic acid analogs described herein are not described with respect to any particular sequence of bases. The present disclosure is directed to divalent nucleobases, compositions comprising the divalent nucleobases, and methods of use of the divalent nucleobases and compounds containing those nucleobases, and the usefulness of any specific embodiments described herein, while depending upon a specific sequence in each instance, is generically applicable. Based on the abundance of published work with nucleic acids, nucleic acid analogs and PNA (e.g., γ PNA), it is expected that any nucleobase sequence attached to the backbone of the described γ PNA oligomers would hybridize in an expected, specific manner with a complementary nucleobase sequence of a target nucleic acid or nucleic acid analog by Watson-Crick or Watson-Crick-like hydrogen bonding. One of ordinary skill would understand that the compositions and methods described herein are sequence-independent and describe novel, generalized compositions comprising divalent nucleobases and related methods.

[0051] In another embodiment, a genetic recognition reagent oligomer is provided, comprising at least one divalent nucleobase. The genetic recognition reagent comprises a backbone and at least two nucleobases, at least one of which is a divalent nucleobase as described herein. This structure is shown schematically in Formula 3:



[0052] where R1 is a backbone monomer residue and R2s are, independently nucleobases, where at least one instance of R2 is a divalent nucleobase, such as any one of JB1-JB16 as shown herein. E are independently end (terminal) groups that are part of the terminal monomer residues, and “n” is any positive integer or 0, for example 48 or less, 28 or less, 23 or less, and 18 or less, including 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18. Typically, all instances of R1 are the same with the exception of the terminal monomer residues which typically have different end-groups E as compared to internal monomers, such as, without limitation NH₂ and C(O)OH or CONH₂ at the respective N-terminal and C-terminal ends for PNAs, and hydroxyl groups at the 5' and 3' ends of nucleic acids.

[0053] In one embodiment of the present invention, the genetic recognition reagents are implemented on an array. Arrays are particularly useful in implementing high-throughput assays, such as genetic detection assays. As used herein, the term "array" refers to reagents, for example the genetic recognition reagents described herein, located at two or more discrete, identifiable and/or addressable locations on a substrate. In one embodiment, an array is an apparatus having two or more discrete, identifiable reaction chambers, such as, without limitation a 96-well dish, in which reactions comprising identified constituents are performed. In an exemplary embodiment, two or more genetic recognition reagents described herein are immobilized onto a substrate in a spatially addressable manner so that each individual primer or probe is located at a different and (addressable) identifiable location on the substrate. Substrates include, without limitation, multi-well plates, silicon chips and beads. In one embodiment, the array comprises two or more sets of beads, with each bead set having an identifiable marker, such as a quantum dot or fluorescent tag, so that the beads are individually identifiable using, for example and without limitation, a flow cytometer. In one embodiment, an array is a multi-well plate containing two or more wells with the described genetic recognition reagents for binding specific sequences. As such, reagents, such as probes and primers may be bound or otherwise deposited onto or into specific locations on an array. Reagents may be in any suitable form, including, without limitation: in solution, dried, lyophilized or glassified. When linked covalently to a substrate, such as an agarose bead or silicon chip, a variety of linking technologies are known for attaching chemical moieties, such as the genetic recognition reagents to such substrates. Linkers and spacers for use in linking nucleic acids, peptide nucleic acids and other nucleic

acid analogs are broadly known in the chemical and array arts and for that reason are not described herein. As a non-limiting example, a γ PNA genetic recognition reagent contains a reactive amine, which can be reacted with carboxyl, cyanogen bromide-, N-hydroxysuccinimide ester-, carbonyldiimidazole- or aldehyde- functional agarose beads, available, for instance from Thermo Fisher Scientific (Pierce Protein Biology Products), Rockford, Illinois and a variety of other sources. The genetic recognition reagents described herein can be attached to a substrate in any manner, with or without linkers. Informatics and/or statistical software or other computer-implemented processes for analyzing array data and/or identifying genetic risk factors from data obtained from a patient sample, are known in the art.

[0054] Certain of the JB divalent compositions are expected to exhibit fluorescence, such as JB1, JB2 and JB3, due to their ring structure. These compositions can, of course, be used as fluorochromes, or the intrinsic fluorescence can be employed as a probe, for example, by binding target sequences in an *in situ* assay or in a gel or blot, such that a target sequence can be visualized.

[0055] Thus, according to one embodiment of the present invention, a method is provided for detection of a target sequence in a nucleic acid, comprising contacting a genetic recognition reagent composition as described herein with a sample comprising nucleic acid and detecting binding of the genetic recognition reagent with a nucleic acid. In one embodiment, the genetic recognition reagent is immobilized on a substrate, for example in an array, and labeled (e.g., fluorescently labeled or radiolabeled) nucleic acid sample is contacted with the immobilized genetic recognition reagent and the amount of labeled nucleic acid specifically bound to the genetic recognition reagent is measured. In a variation, genetic recognition reagent or a nucleic acid comprising a target sequence of the genetic recognition reagent is bound to a substrate, and a labeled nucleic acid comprising a target sequence of the genetic recognition reagent or a labeled genetic recognition reagent is bound to the immobilized genetic recognition reagent or nucleic acid, respectively to form a complex. In one embodiment, the nucleic acid of the complex comprises a partial target sequence so that a nucleic acid comprising the full target sequence would out-compete the complexed nucleic acid for the genetic recognition reagent. The complex is then exposed to a nucleic acid sample and loss of bound label from the complex could be detected and quantified according to standard methods, facilitating quantification of a nucleic acid marker in the nucleic acid sample. These are merely two of a large number of possible analytical assays that can be used to detect or quantify the presence of a specific nucleic acid in a nucleic acid sample.

[0056] By “immobilized” in reference to a composition such as a nucleic acid or genetic recognition reagent as described herein, it is meant attached to a substrate of any physical structure or chemical composition. The immobilized composition is immobilized by any method useful in the context of the end use. The composition is immobilized by covalent or non-covalent methods, such as by covalent linkage of amine groups to a linker or spacer, or by non-covalent bonding, including Van derWaals and/or hydrogen bonding. A “label” is a chemical moiety that is useful in detection of, or purification of a molecule or composition comprising the label. A label may be, for example and without limitation, a radioactive moiety, such as ^{14}C , ^{32}P , ^{35}S , a fluorescent dye, such as fluorescein isothiocyanate or a cyanine dye, an enzyme, or a ligand for binding other compounds such as biotin for binding streptavidin, or an epitope for binding an antibody. A multitude of such labels, and methods of use thereof are known to those of ordinary skill in the immunology and molecular biology arts.

[0057] In yet another embodiment of the present invention, a method of isolation and purification of a nucleic acid containing a target sequence is provided. In one non-limiting embodiment, a genetic recognition reagent as described herein is immobilized on a substrate, such as a bead (for example and without limitation, an agarose bead, a bead containing a fluorescent marker for sorting, or a magnetic bead), porous matrix, surface, tube, *etc.* A nucleic acid sample is contacted with the immobilized genetic recognition reagent and nucleic acids containing the target sequence bind to the genetic recognition reagent. The bound nucleic acid is then washed to remove unbound nucleic acids, and the bound nucleic acid is then eluted, and can be precipitated or otherwise concentrated by any useful method as are broadly known in the molecular biological arts.

[0058] In a further embodiment, kits are provided. A kit comprises at a minimum a vessel of any form, including cartridges for automated PNA synthesis, which may comprise one or more vessels in the form of compartments. Vessels may be single-use, or contain sufficient contents for multiple uses. A kit also may comprise an array. A kit may optionally comprise one or more additional reagents for use in making or using genetic recognition reagents in any embodiment described herein. The kit comprises a vessel containing any divalent nucleobase described herein, or monomers or genetic recognition reagents according to any embodiment described herein. Different nucleobases, monomers or genetic recognition reagents are typically packaged into separate vessels, which may be separate compartments in a cartridge.

Examples

[0059] The examples described herein illustrate advantages of the disclosed technology over traditional oligonucleotide systems. γ PNA containing JB 1-4 as recognition elements (γ PNAJB1-4) can be designed to bind to any sequence of double-stranded DNA (dsDNA) or RNA (dsRNA) under physiological conditions through direct hydrogen bonding interactions. These probes typically are as short 3nt in length and as long as 25nt in length. The shorter probes, 10nt or less, are especially easier to synthesize and scale up, via solution - as opposed to solid-phase chemistry, and are more readily taken up by cells.

[0060] Recognition of dsDNA or dsRNA by γ PNAJB1-4 is more sequence-specific than the existing antigene or antisense platforms because a mismatch on one face of JB would result in a mismatch on the second face. Binding of γ PNAJB1-4 to DNA or RNA is more selective for double-strand than for single-strand or primary (unstructured) sequences. Additionally, γ PNA containing JB 1-16 (γ PNAJB1-16) can be designed to bind to any secondary or tertiary structure of RNA and modulate its folding patterns and three-dimensional architectures.

[0061] Furthermore, these examples also provide distinctive features and advantages over small-molecule drugs or other ligands. Recognition of the γ PNAJB probes occurs in a sequence-specific and predictable manner in accordance with the Watson-Crick and Janus base-pairing rules. Because they are modular in nature, γ PNAJB probes could be modified to match the sequence of any RNA. This is an essential requirement for countering drug resistance due to emergence of genetic mutations—in particular, for genetic targets, such as that of cancer, bacteria, viral and parasite genomes or transcriptomes.

[0062] Example applications include, but are not limited to, reagents for detection, analysis and purification of DNA and RNA molecules, and for molecular tools for regulating gene expression, correcting genetic mutations, modulating RNA splicing, controlling microRNA functions, and modulation of RNA folding patterns and architectures. These applications will positively contribute to diagnostics; for example to detect DNA and RNA sequence and structural variations. Additionally this platform will advance medicine; as an example for use as therapeutics for treating a variety of genetic diseases associated with misregulation of gene expression including cancer, deregulation of RNA splicing, DNA and RNA unstable repeats, and fungal, bacterial and viral biology and infections.

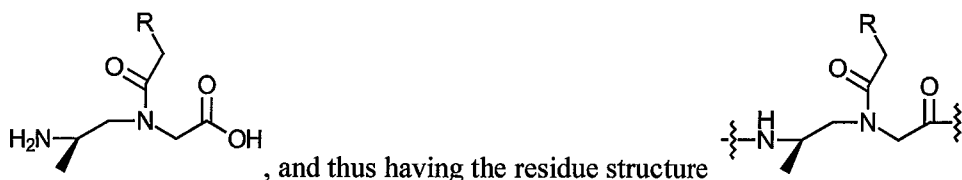
[0063] The applications of this nucleic acid platform using γ PNA and the Janus nucleobases span basic research in biology and biotechnology, diagnostics, and therapeutics. This platform, which targets double-stranded DNA and RNA has been demonstrated and described

in accordance with several examples, which are intended to be illustrative in all aspects rather than restrictive. Thus, the present invention is capable of many variations in detailed implementation, which may be derived from the description contained herein by a person of ordinary skill in the art.

Example 1 – Computer simulation of interactions of γ PNA with DNA and RNA

[0064] DNA binding properties of γ PNA include efficiencies from backbone preorganization (as defined by the stereochemistry at the γ -position, which determines the helical sense of the oligomer) and improvements in hydrogen-bonding and base-stacking capabilities. Divalent nucleobases JB 1-4, were designed to bind to the perfectly matched, and JB 5-16 were designed to bind to mismatched base-pairs in the DNA or RNA double helix. In both cases binding occurs via strand invasion, whereby one face of the divalent nucleobase forms Watson-Crick-like hydrogen-bonding interactions with natural bases on one strand of DNA or RNA double helix, while a second face forms ‘Janus’ hydrogen-bonding interactions with the complementary strand of the DNA or RNA target. As indicated above, Figure 1 illustrates these hydrogen-bonding interactions, while comparing to Watson-Crick-like hydrogen-bonding interactions with natural bases.

[0065] In one embodiment, the overall geometry, bond distance, and chemical functionalities of these divalent “Janus” nucleobases are carefully crafted so that they can be used interchangeably to target the perfectly-matched as well as mismatched binding sites, or a combination thereof. It is demonstrated from these examples that this platform of γ PNA monomers/oligomers containing the JBs are more selective for double-stranded DNA or RNA, in addition to recognizing secondary and tertiary structures. DNA- γ PNA interactions were modeled by computer simulation using AMBER10 (University of California, San Francisco, CA). The γ PNA is based on alanine, having the monomer structure



[0066] Figure 5 illustrates the simulated structures of γ PNAJB1-4 invading DNA double helix.

[0067] Figure 6 illustrates the structures of γ PNAJB1-16 invading an RNA secondary structure containing perfectly-matched and mismatched base-pairs.

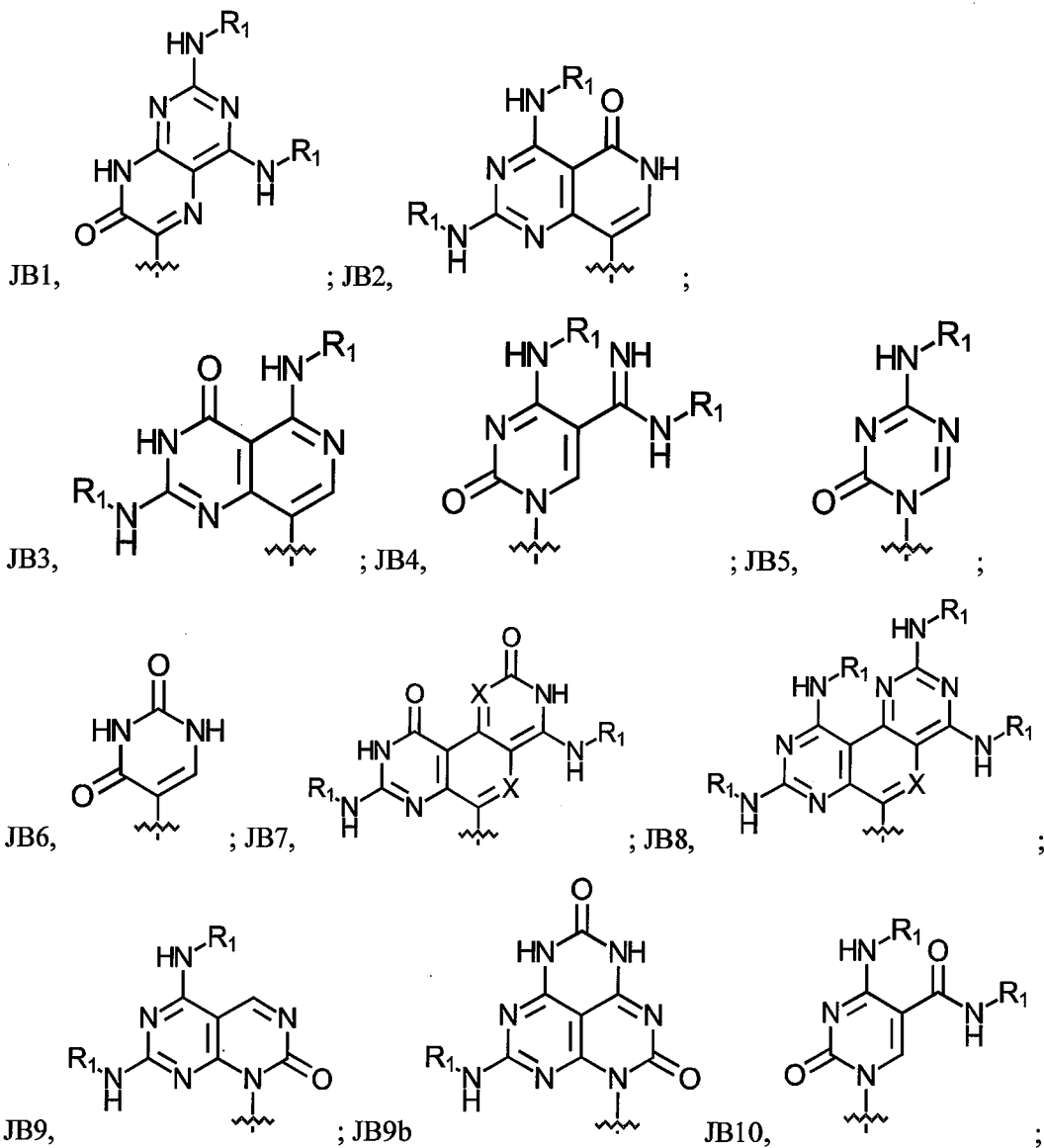
Example 2 – Synthesis of divalent nucleobases, nucleotide monomers and γ PNA oligomers

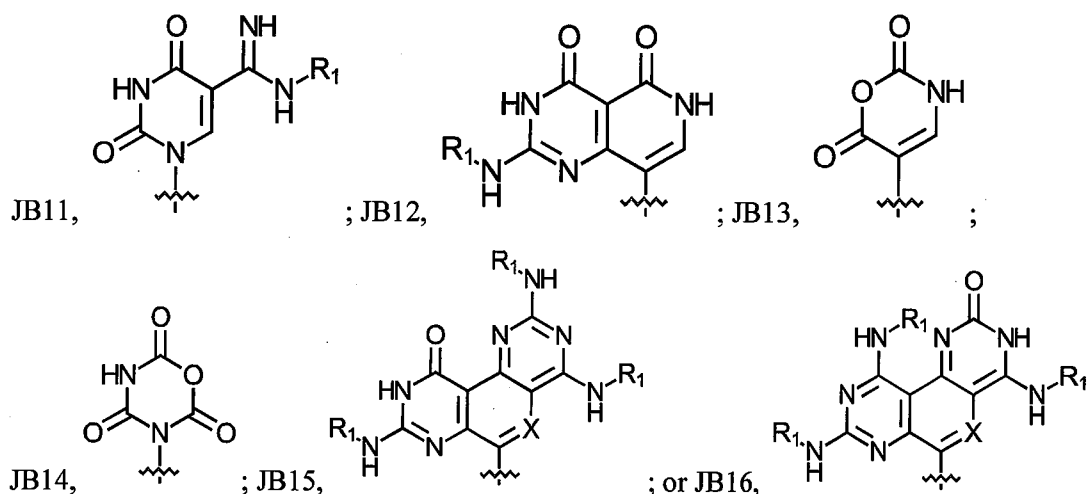
[0068] Examples of reactions performed to synthesis JBs 1-16 are shown in Figure 7. Upon synthesis of these specialized bases, the corresponding monomers can be synthesized and an example synthesis scheme is described in Figure 8. Figure 9 illustrates the synthesis of oligomers containing the γ PNA oligomers.

[0069] The present invention has been described with reference to certain exemplary embodiments, dispersible compositions and uses thereof. However, it will be recognized by those of ordinary skill in the art that various substitutions, modifications or combinations of any of the exemplary embodiments may be made without departing from the spirit and scope of the invention. Thus, the invention is not limited by the description of the exemplary embodiments, but rather by the appended claims as originally filed.

We claim:

1. A genetic recognition reagent comprising a plurality of nucleobase residues attached to a nucleic acid or nucleic acid analog backbone, in which at least one nucleobase is a divalent nucleobase chosen from the following:

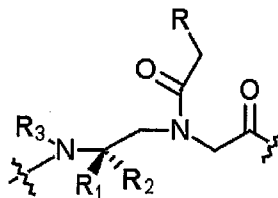




in which each instance of R₁ is, independently, a protecting group or H and X is CH or N.

2. The genetic recognition reagent of claim 1, in which each instance of R is, independently, H or a protecting group chosen from: methyl, formyl, ethyl, acetyl, anisyl, benzyl, benzoyl, carbamate, trifluoroacetyl, diphenylmethyl, triphenylmethyl, N-hydroxysuccinimide, benzyloxymethyl, benzyloxycarbonyl, 2-nitrobenzoyl, *t*-Boc (*tert*-butyloxycarbonyl), 4-methylbenzyl, 4-nitrophenyl, 2-chlorobenzyloxycarbonyl, 2-bromobenzyloxycarbonyl, 2,4,5-trichlorophenyl, thioanisyl, thiocresyl, cbz (carbobenzyloxy), *p*-methoxybenzyl carbonyl, 9-fluorenylmethyloxycarbonyl, pentafluorophenyl, *p*-methoxybenzyl, 3,4-dimethoxybenzyl, *p*-methoxyphenyl, 4-toluenesulfonyl, *p*-nitrobenzenesulfonates, 9-fluorenylmethyloxycarbonyl, 2-nitrophenylsulfonyl, 2,2,5,7,8-pentamethyl-chroman-6-sulfonyl, and *p*-bromobenzenesulfonyl.
3. The genetic recognition reagent of claim 1, in which the backbone is chosen from one of a DNA, RNA, peptide nucleic acid (PNA), phosphorothioate, locked nucleic acid, unlocked nucleic acid, 2'-O-methyl-substituted RNA, morpholino nucleic acid, threose nucleic acid, or glycol nucleic acid backbone, or any combination thereof.
4. The genetic recognition reagent of claim 1, in which the backbone is a peptide nucleic acid (PNA) backbone.
5. The genetic recognition reagent of claim 1, in which the backbone is a gamma peptide nucleic acid (γ PNA) backbone.
6. The genetic recognition reagent of claim 5, in which the backbone is PEGylated, with one or more PEG moieties of two to fifty (-O-CH₂-CH₂-) residues.

7. The genetic recognition reagent of claim 1, in which the backbone is a γ PNA

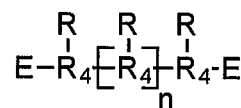


backbone in which the backbone monomer residue is $\text{---N(R}_3\text{)---CH(R}_1\text{)---CH}_2\text{---C(=O)---N(R)---C(=O)---}$, where R1, R2 and R3 are, independently, H, amino acid side chains, linear or branched (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₁-C₈)hydroxyalkyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene, (C₃-C₈)cycloalkyl(C₁-C₆)alkylene, PEGylated moieties of the preceding comprising from 1 to 50 (-O-CH₂-CH₂-) residues, -CH₂-(OCH₂-CH₂)_qOP₁, -CH₂-(OCH₂-CH₂)_q-NHP₁, -CH₂-(OCH₂-CH₂-O)_q-SP₁ and -CH₂-(SCH₂-CH₂)_q-SP₁, -CH₂-(OCH₂-CH₂)_r-OH, -CH₂-(OCH₂-CH₂)_r-NH₂, -CH₂-(OCH₂-CH₂)_r-NHC(NH)NH₂, or -CH₂-(OCH₂-CH₂)_r-S-S[CH₂CH₂]_sNHC(NH)NH₂, where P₁ is selected from the group consisting of H, (C₁-C₈)alkyl, (C₂-C₈)alkenyl, (C₂-C₈)alkynyl, (C₃-C₈)aryl, (C₃-C₈)cycloalkyl, (C₃-C₈)aryl(C₁-C₆)alkylene and (C₃-C₈)cycloalkyl(C₁-C₆)alkylene; q is an integer from 0 to 10, inclusive; r and s are each independently integers from 1 to 50, inclusive; where R1 and R2 are different and one of R1 or R2 is H.

8. The genetic recognition reagent of claim 7, in which R2 is H, R1 is an amino acid side chain that is optionally PEGylated, with one or more PEG moieties of one to twelve (-O-CH₂-CH₂-) residues.

9. The genetic recognition reagent of claim 1, in which the nucleobases are arranged in a sequence complementary to a target sequence of a nucleic acid.

10. The genetic recognition reagent of claim 1, having from 3 to 25 nucleobases.



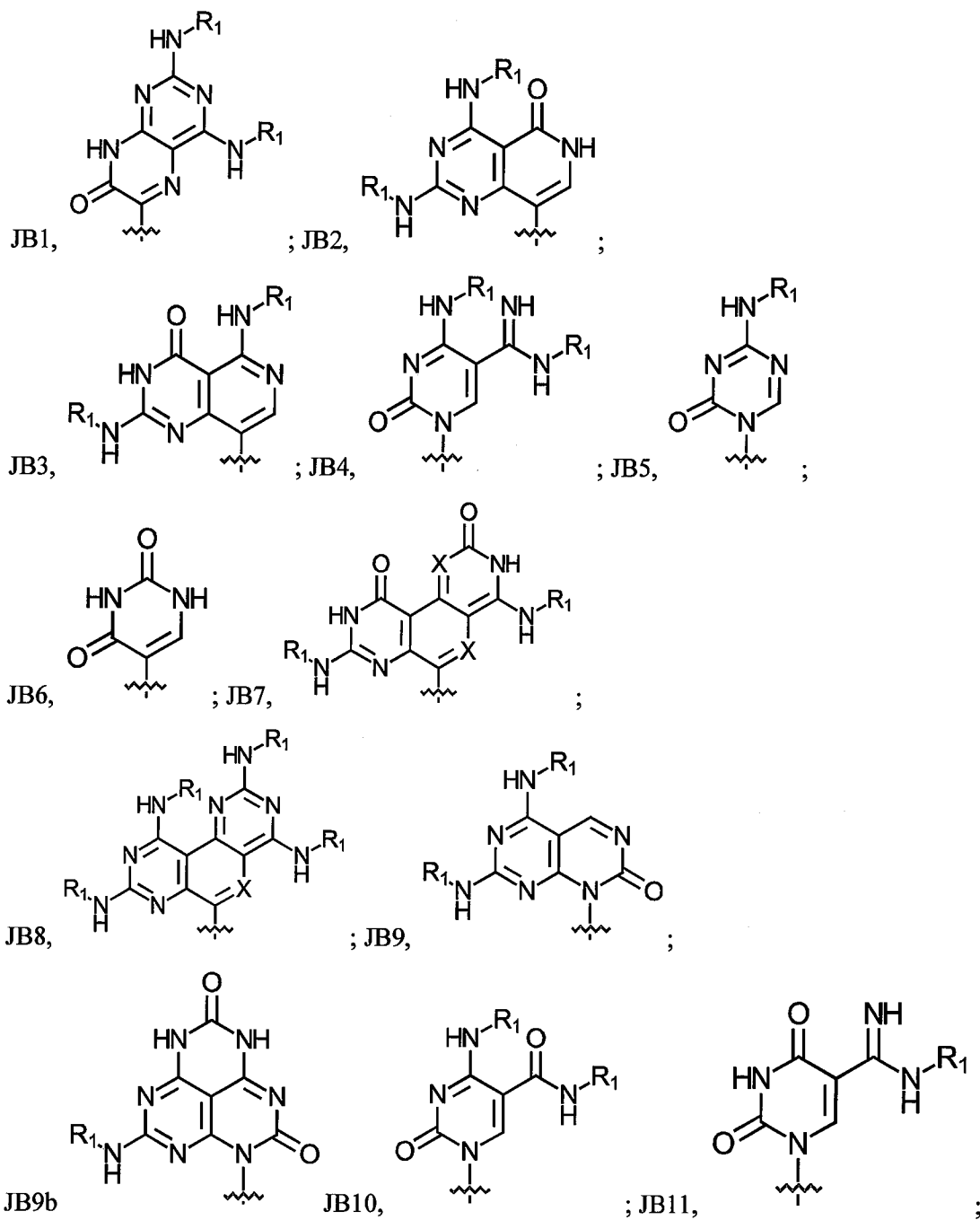
11. The genetic recognition reagent of claim 1, having the structure:

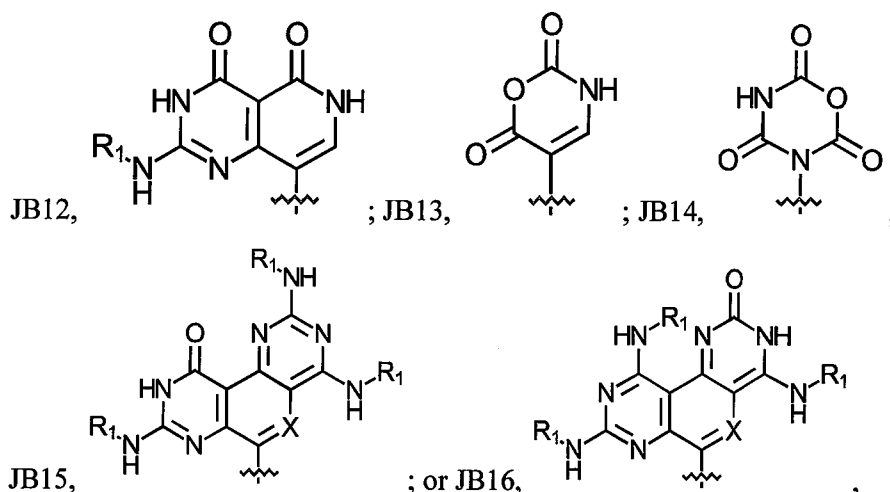
where each instance of R4 is a backbone monomer residue and each instance of R is a nucleobase, where at least one instance of R the divalent nucleobase, E are independently end groups, and "n" is zero or a positive integer ranging from 1 to 48, in which the sequence of nucleobases R is complementary to a target sequence of a nucleic acid.

12. The genetic recognition reagent of claim 1, in which the divalent nucleobases are, independently chosen from JB1, JB2, JB3 and JB4.

13. The genetic recognition reagent of claim 1, in which all instances of R1 are H.

14. A monomer for production of a genetic recognition reagent comprising a backbone monomer for a genetic recognition reagent covalently attached to a divalent nucleobase nucleobase chosen from the following:





in which each instance of R1 is, independently, a protecting group or H and X is CH or N.

15. The monomer of claim 14, in which the nucleobase is one of JB1, JB2, JB3 or JB4.

16. The monomer of claim 14, in which R1 is a protecting group.

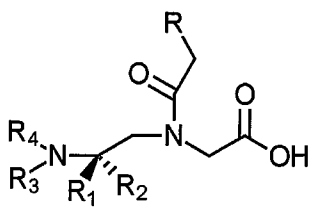
17. The monomer of claim 16, in which the protecting group is chosen from one or more of: methyl, formyl, ethyl, acetyl, anisyl, benzyl, benzoyl, carbamate, trifluoroacetyl, diphenylmethyl, triphenylmethyl, N-hydroxysuccinimide, benzyloxymethyl, benzyloxycarbonyl, 2-nitrobenzoyl, *t*-Boc (*tert*-butyloxycarbonyl), 4-methylbenzyl, 4-nitrophenyl, 2-chlorobenzyloxycarbonyl, 2-bromobenzyloxycarbonyl, 2,4,5-trichlorophenyl, thioanisyl, thiocresyl, cbz (carbobenzyloxy), *p*-methoxybenzyl carbonyl, 9-fluorenylmethyloxycarbonyl, pentafluorophenyl, *p*-methoxybenzyl, 3,4-dimethoxybenzyl, *p*-methoxyphenyl, 4-toluenesulfonyl, *p*-nitrobenzenesulfonates, 9-fluorenylmethyloxycarbonyl, 2-nitrophenylsulfonyl, 2,2,5,7,8-pentamethyl-chroman-6-sulfonyl, and *p*-bromobenzenesulfonyl.

18. The monomer of claim 14, in which the monomer is a peptide nucleic acid (PNA) monomer.

19. The monomer of claim 14, in which the monomer is a gamma peptide nucleic acid (γ PNA) backbone.

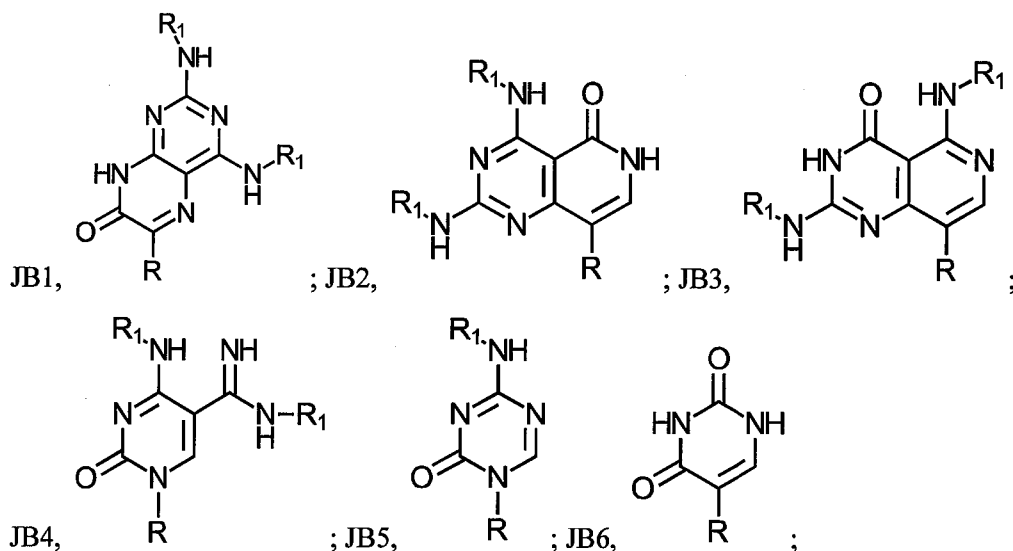
20. The monomer of claim 14, in which the backbone monomer is PEGylated, with one or more PEG moieties of two to fifty (-O-CH₂-CH₂-) residues.

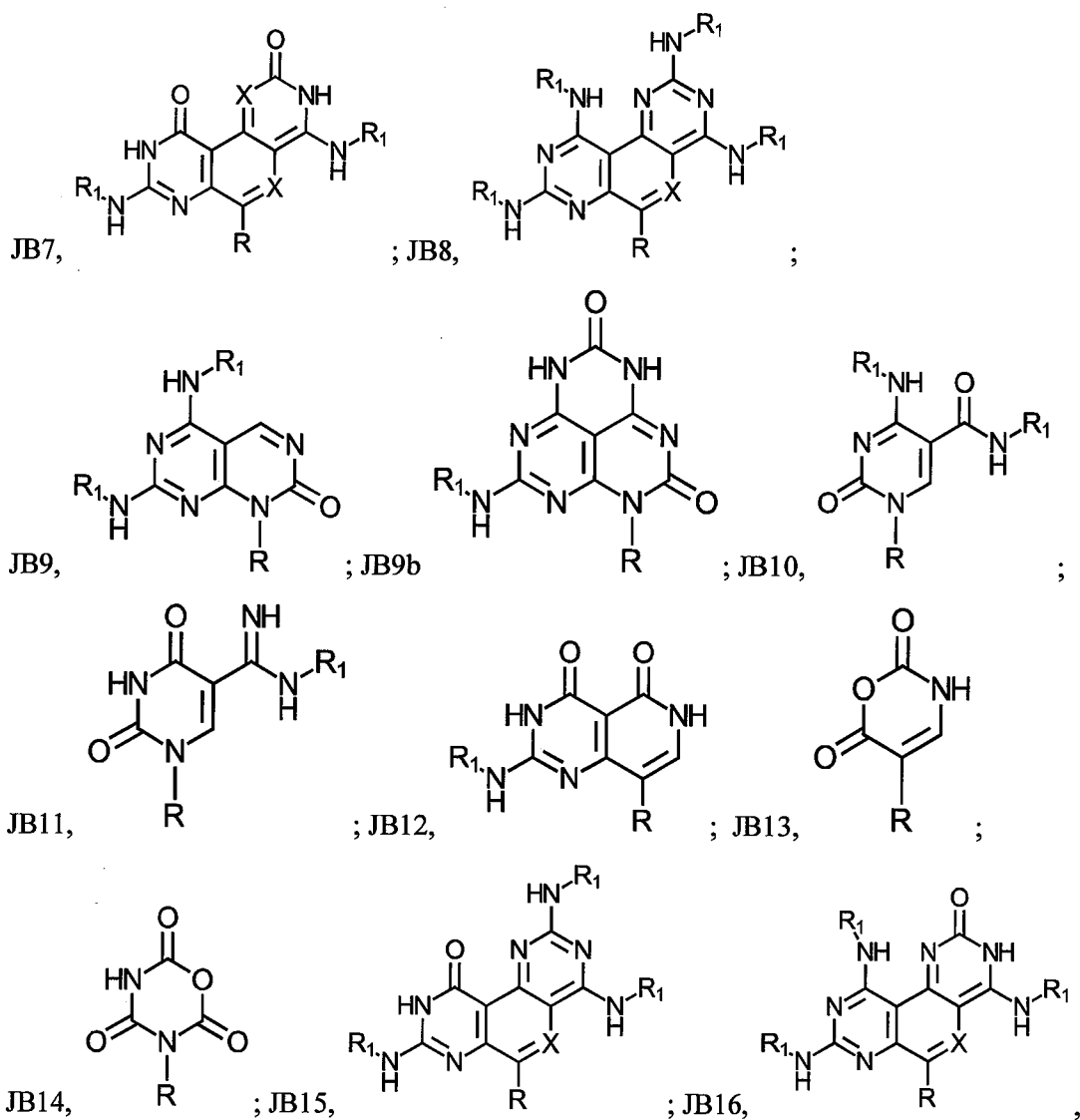
21. The monomer of claim 14, in which the monomer is a γ PNA monomer having the

structure:  , where R1, R2 and R4 are, independently, H, amino acid side chains, linear or branched (C1-C8)alkyl, (C2-C8)alkenyl, (C2-C8)alkynyl, (C1-C8)hydroxyalkyl, (C3-C8)aryl, (C3-C8)cycloalkyl, (C3-C8)aryl(C1-C6)alkylene, (C3-C8)cycloalkyl(C1-C6)alkylene, PEGylated moieties of the preceding comprising from 1 to 50 (-O-CH2-CH2-) residues, -CH2-(OCH2-CH2)qOP1, -CH2-(OCH2-CH2)q-NHP1, -CH2-(OCH2-CH2)q-SP1 and -CH2-(SCH2-CH2)q-SP1, -CH2-(OCH2-CH2)r-OH, -CH2-(OCH2-CH2)r-NH2, -CH2-(OCH2-CH2)r-NHC(NH)NH2, or -CH2-(OCH2-CH2)r-S-S[CH2CH2]sNHC(NH)NH2, where P1 is selected from the group consisting of H, (C1-C8)alkyl, (C2-C8)alkenyl, (C2-C8)alkynyl, (C3-C8)aryl, (C3-C8)cycloalkyl, (C3-C8)aryl(C1-C6)alkylene and (C3-C8)cycloalkyl(C1-C6)alkylene; q is an integer from 0 to 10, inclusive; r and s are each independently integers from 1 to 50, inclusive; where R1 and R2 are different, and one of R1 or R2 is H, and R3 is H or a protecting group.

22. The monomer of claim 21, in which R2 is H, R1 is an amino acid side chain that is optionally PEGylated, with one or more PEG moieties of one to twelve (-O-CH2-CH2-) residues, and R3 is a protecting group.

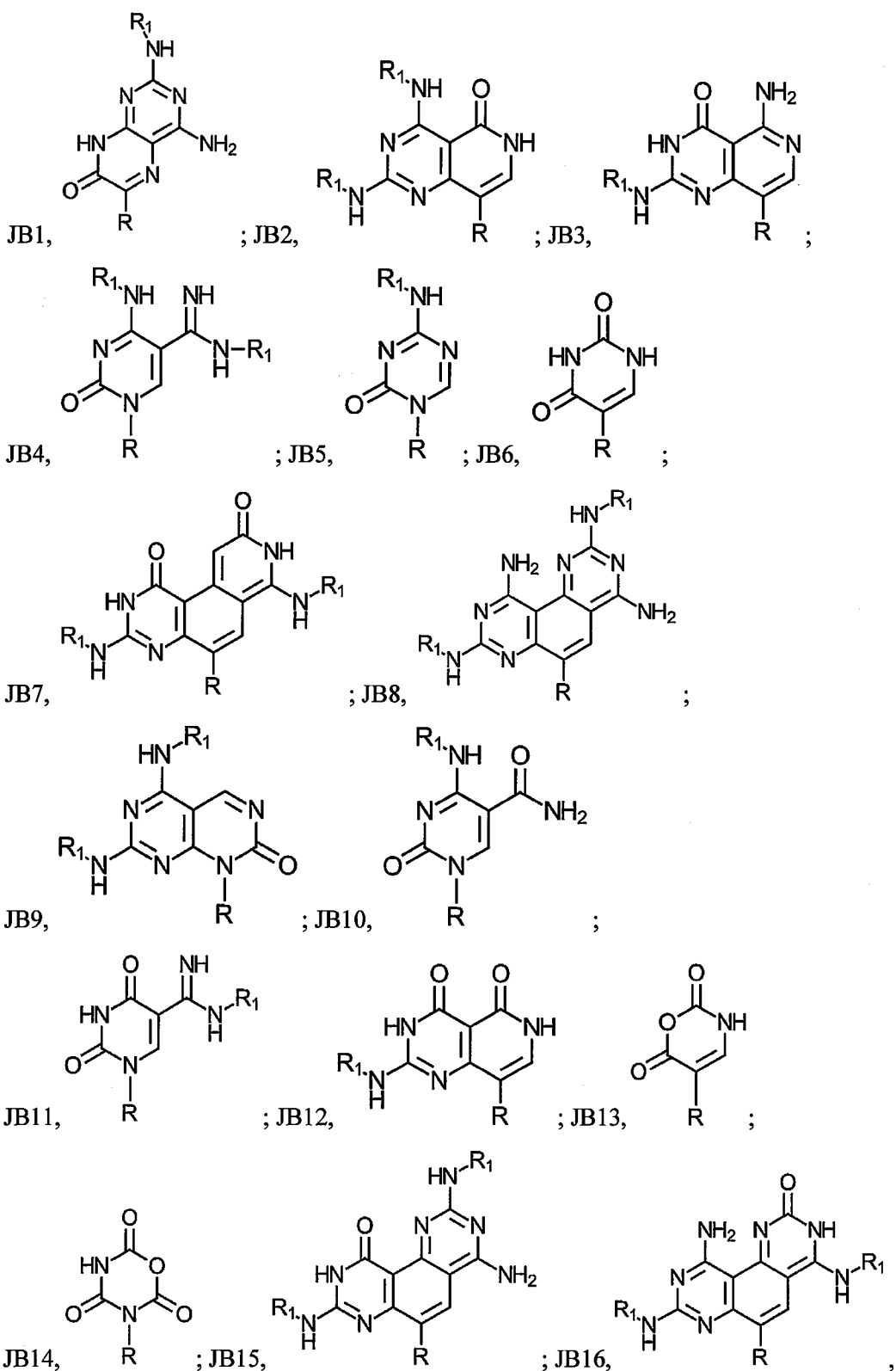
23. A nucleobase chosen from one the following:





in which each instance of R₁ is, independently a protecting group or H, R is a reactive group and X is CH or N.

24. The nucleobase of claim 23, chosen from one the following:

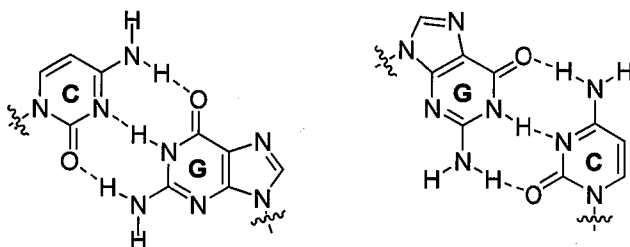
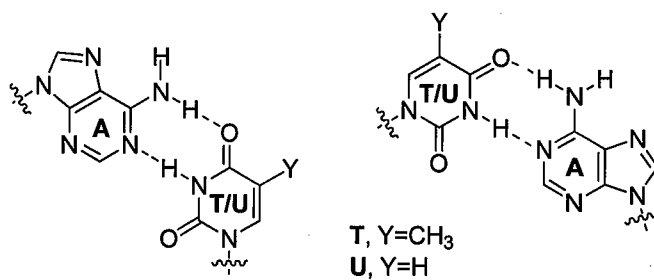
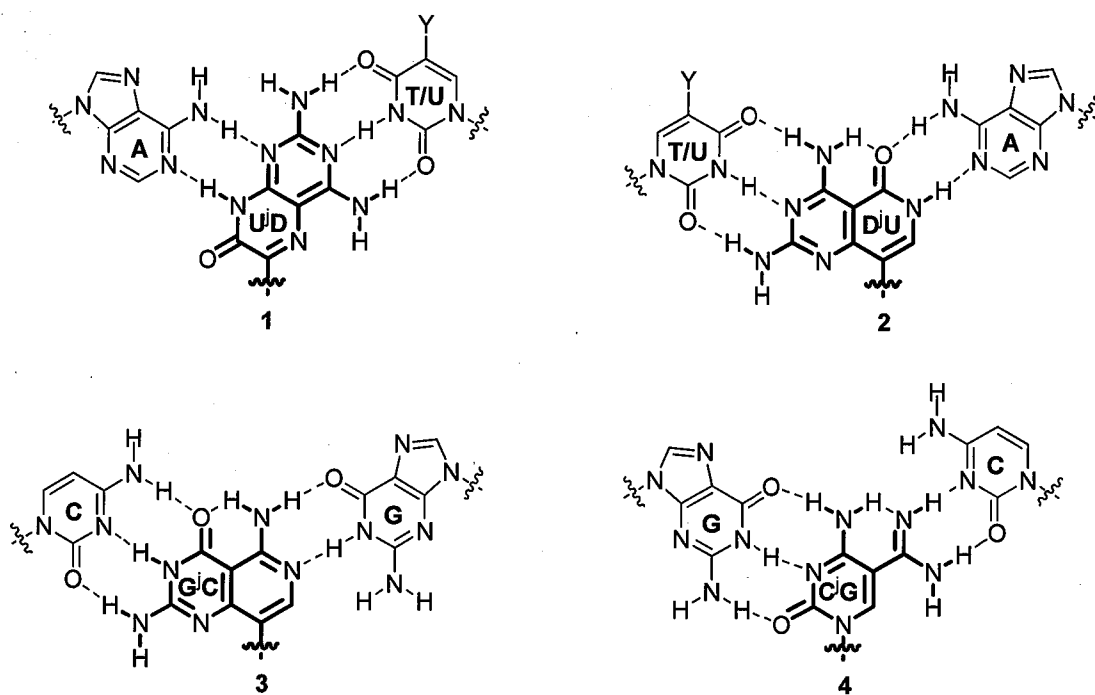


in which each instance of R1 is, independently a protecting group or H and R is carboxyl.

25. A kit comprising a monomer of any one of claims 14-22 in a vessel.

26. The kit of claim 23, further comprising monomers comprising each of JB1, JB2, JB3, JB4, JB5, JB6, JB7, JB8, JB9, JB9b, JB10, JB11, JB12, JB13, JB14, JB15, JB16 nucleobases separate vessels.
27. An array comprising a genetic recognition reagent of any one of claims 1- 13.
28. A method of detection of a target sequence in a nucleic acid, comprising contacting a genetic recognition reagent of any one of claims 1-13 with a sample comprising nucleic acid and detecting binding of the genetic recognition reagent with a nucleic acid.
29. A method of isolation and purification of a nucleic acid containing a target sequence, comprising, contacting a nucleic acid sample with a genetic recognition reagent of any of claims 1-13, separating the nucleic acid sample from the genetic recognition reagent, leaving any nucleic acid bound to the genetic recognition reagent bound to the genetic recognition reagent, and separating the genetic recognition reagent from any nucleic acid bound to the genetic recognition reagent.
30. The method of claim 29 in which the genetic recognition reagent is immobilized on a substrate, comprising contacting a nucleic acid with the substrate, washing the substrate to remove unbound nucleic acid from the substrate, but leaving bound nucleic acid bound to the substrate, and eluting the bound nucleic acid from the substrate.

1/20

**Fig. 1A****Fig. 1B**

2/20

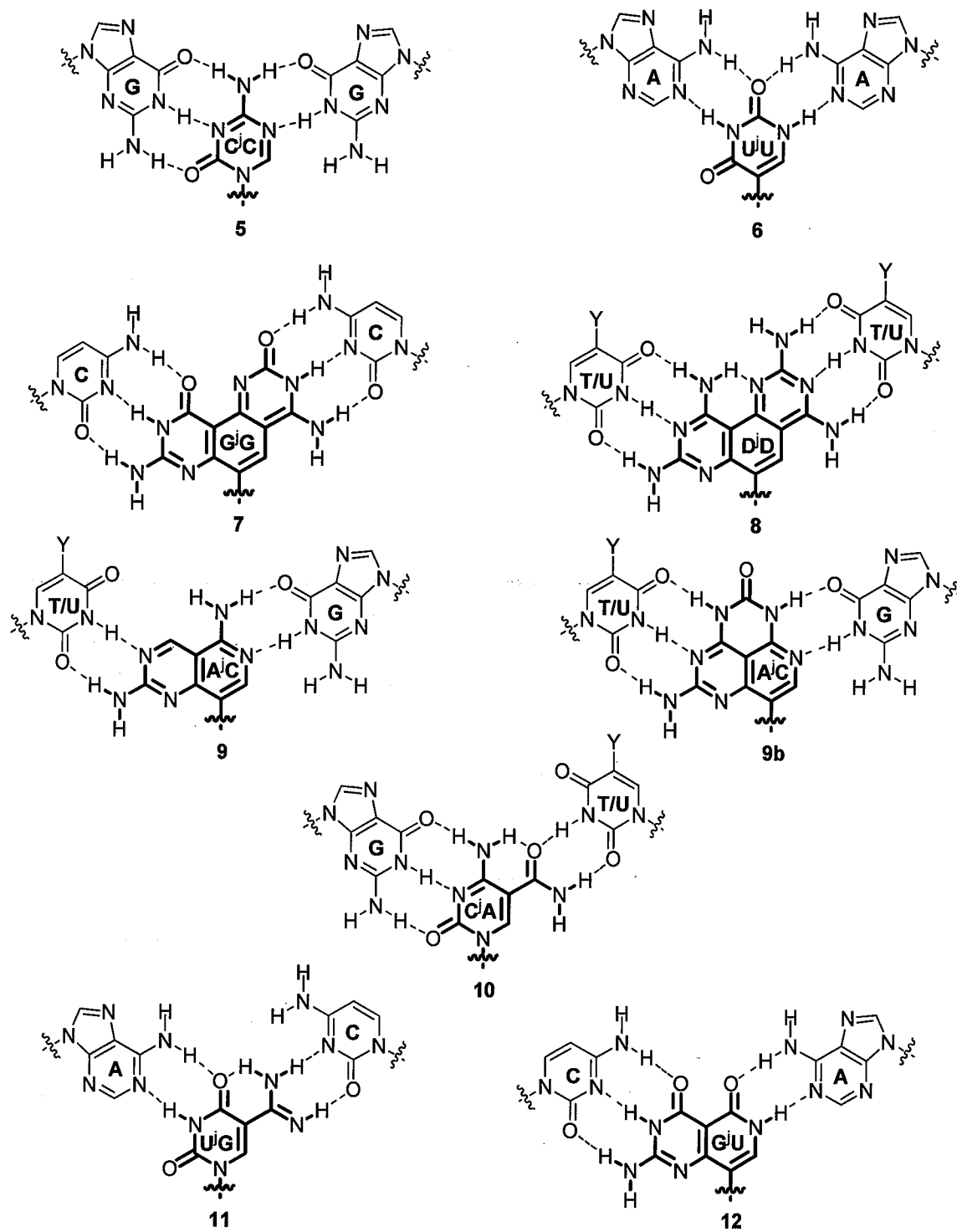
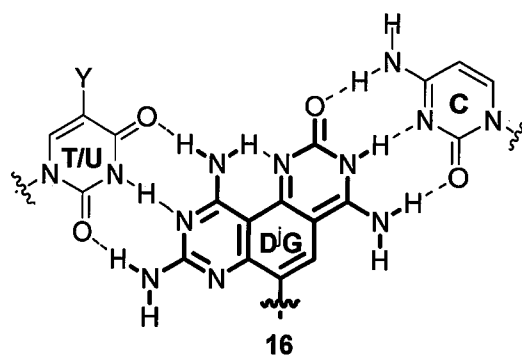
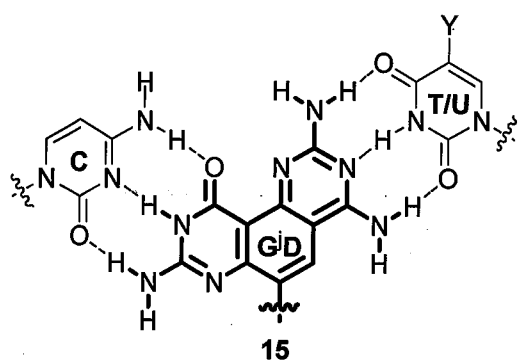
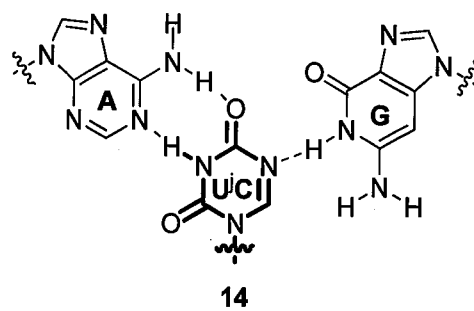
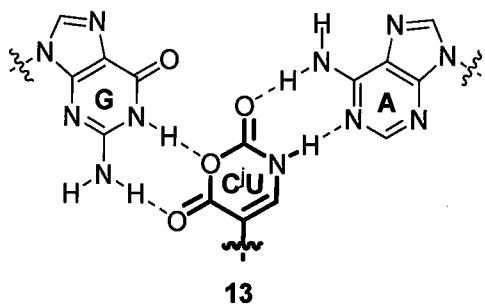
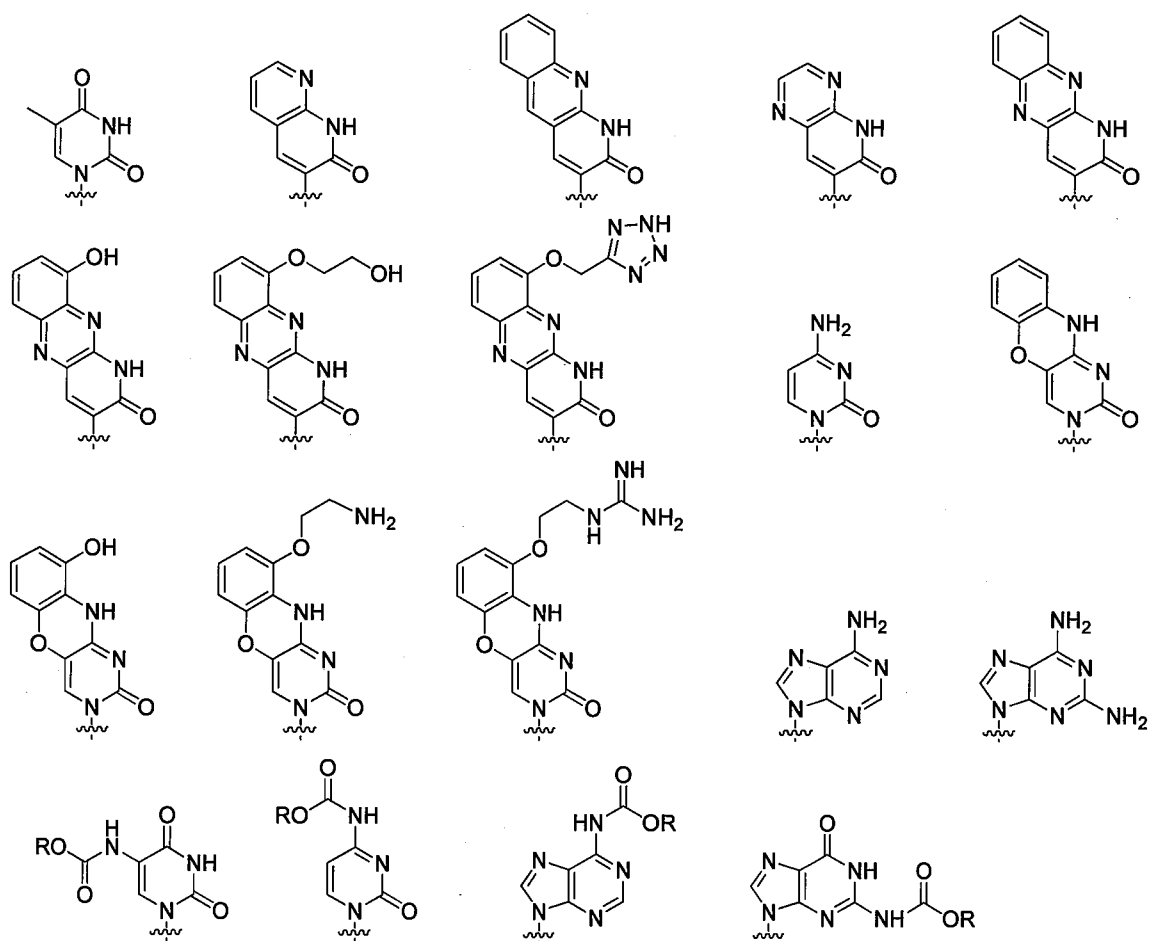


Fig. 1C

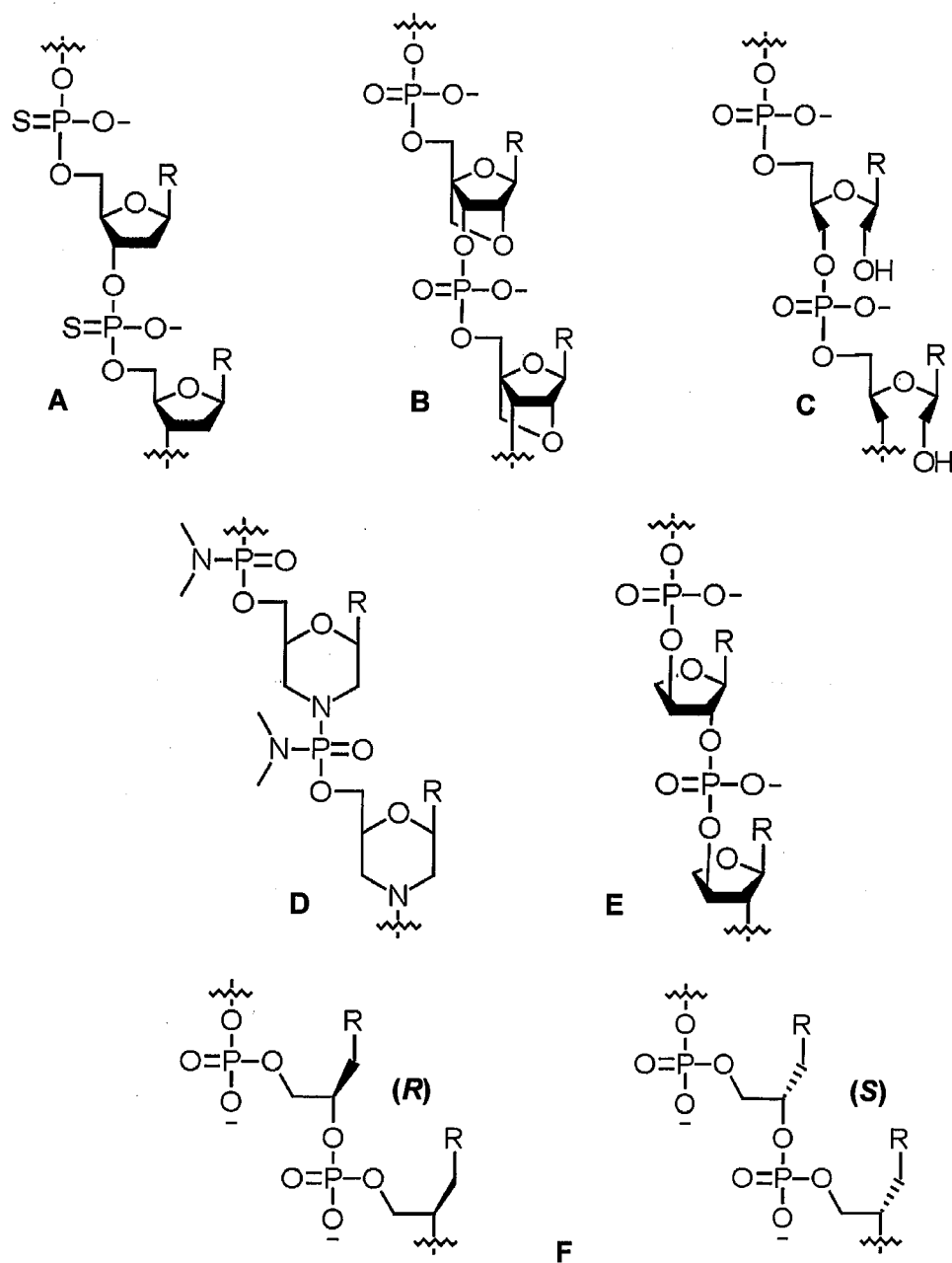
3/20

**Fig 1C (continued)**

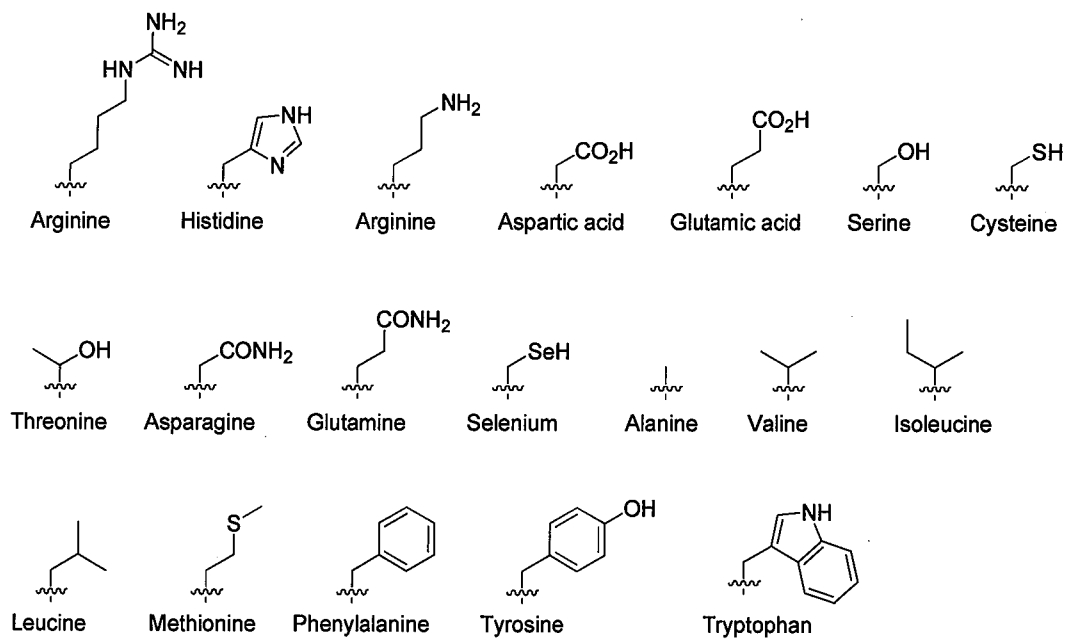
4/20

**Fig. 2**

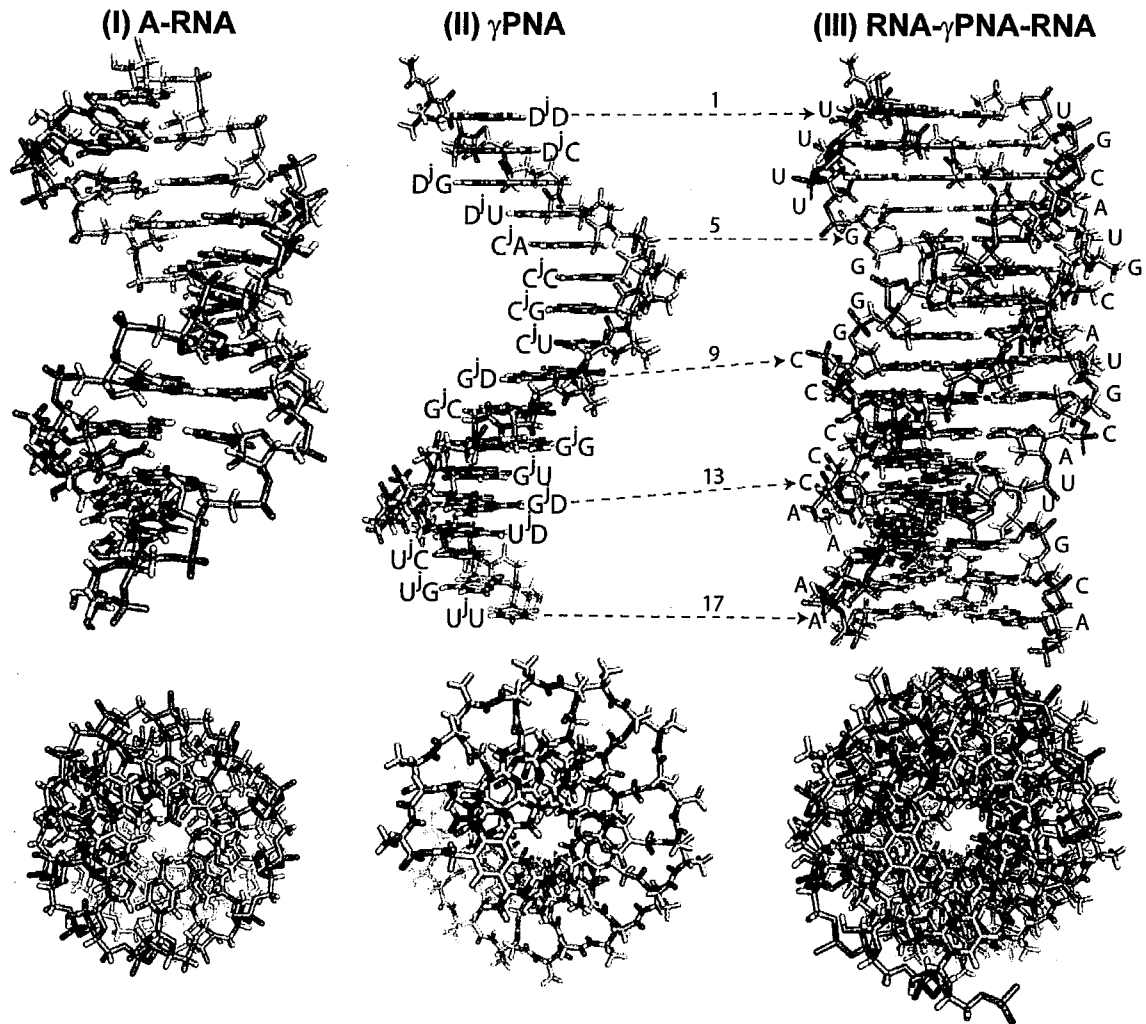
5/20

**Fig. 3**

6/20

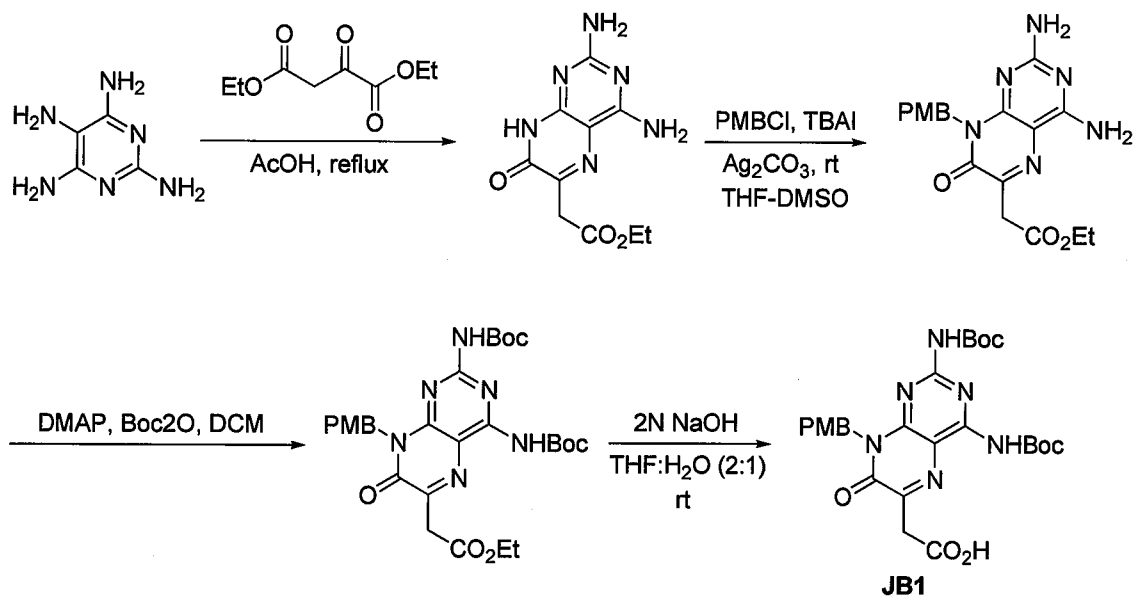
**Fig. 4**

8/20

**Fig. 6**

9/20

JB1



JB2

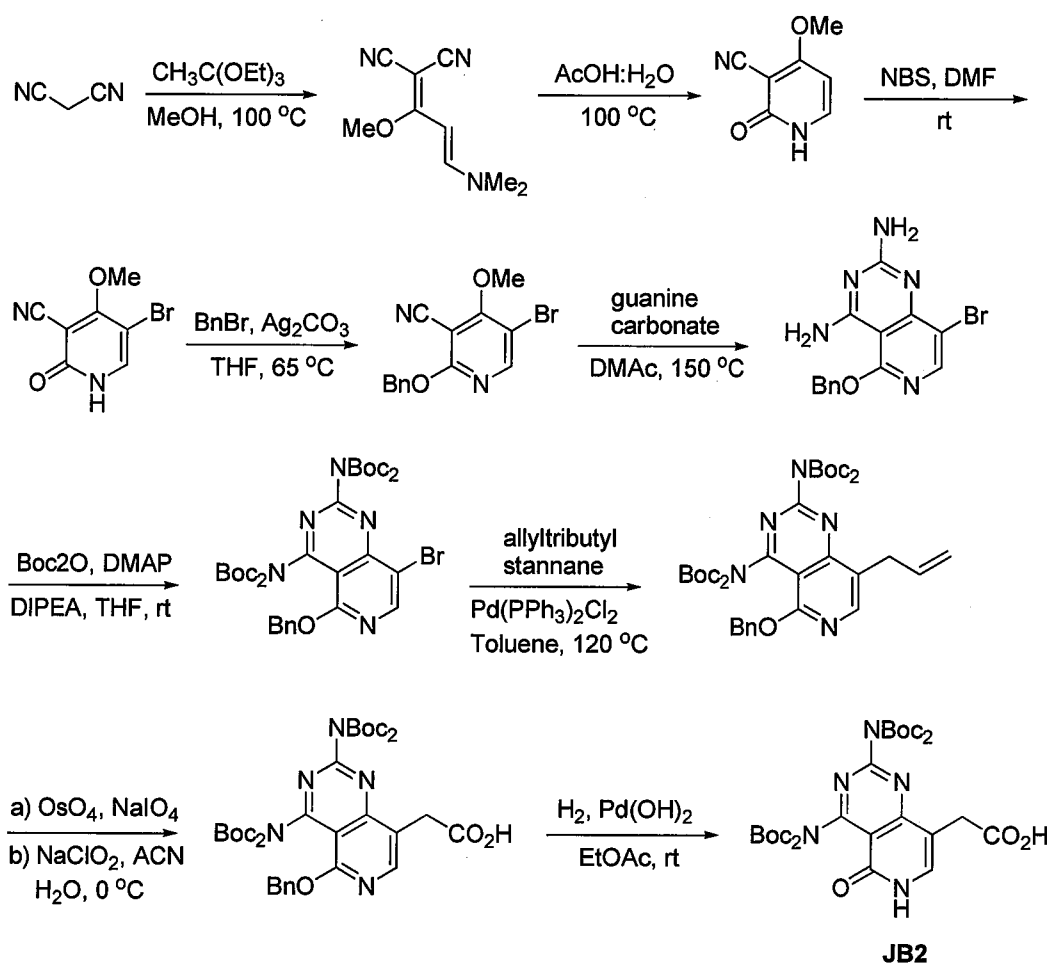


Fig. 7

10/20

JB3

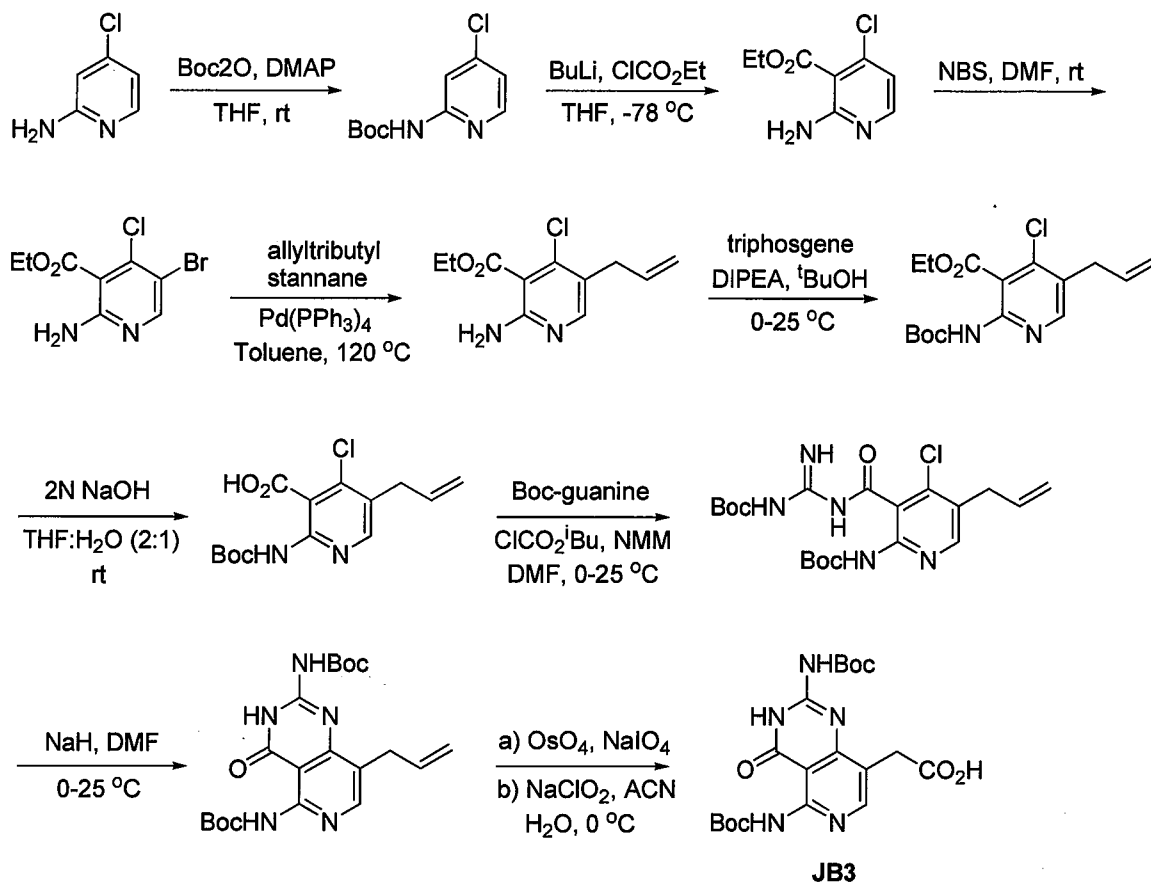


Fig. 7 (continued)

11/20

JB4

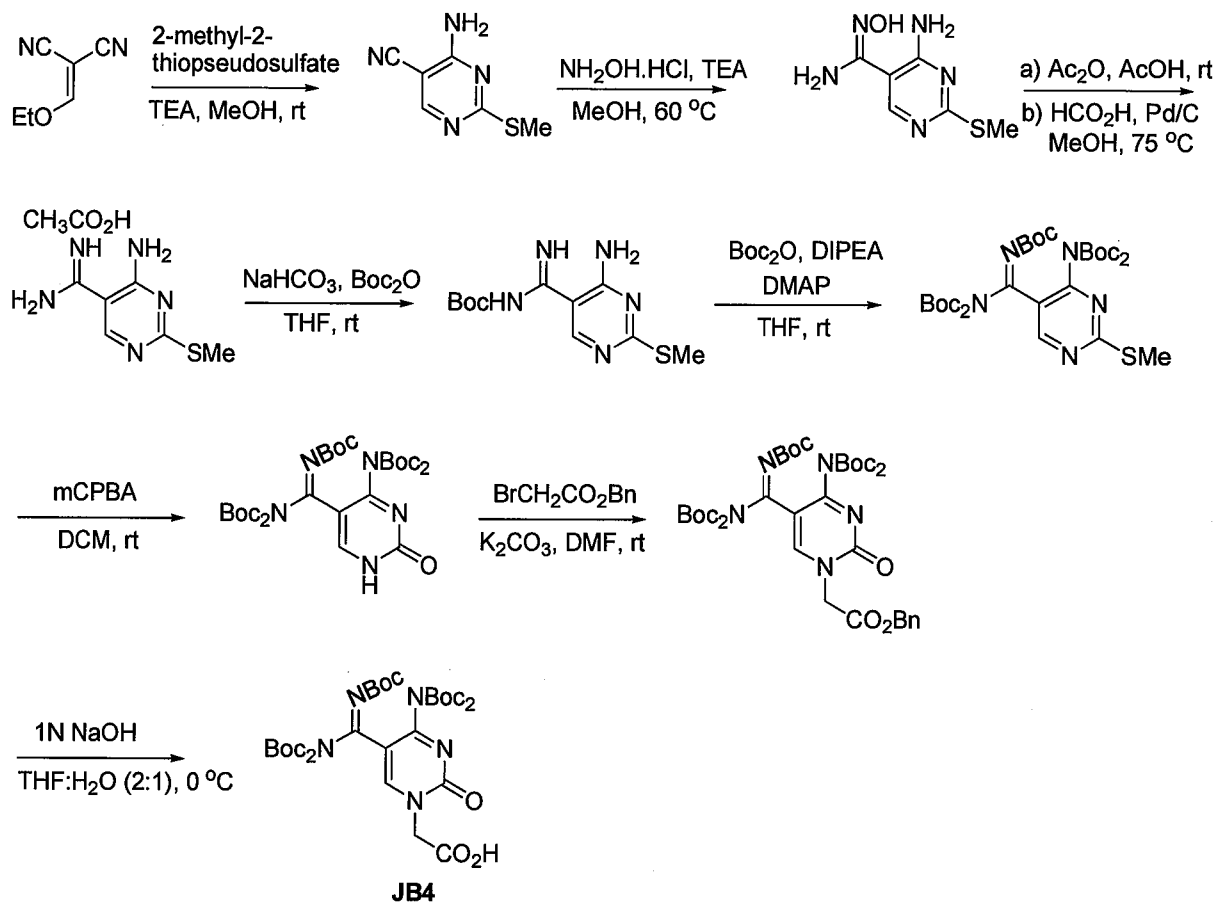
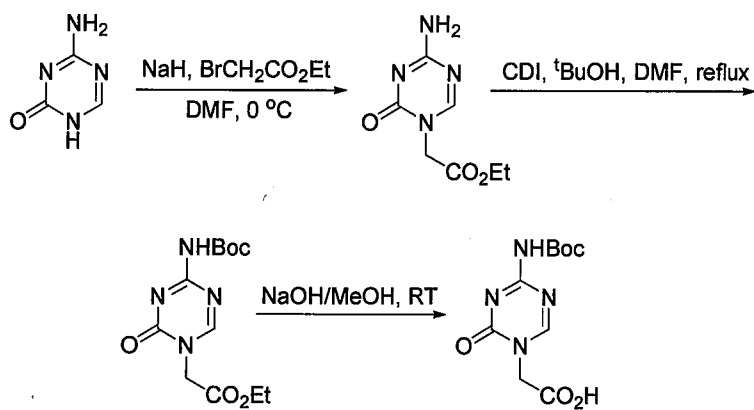


Fig. 7 (continued)

12/20

JB5

**Fig. 7 (continued)**

13/20

JB7

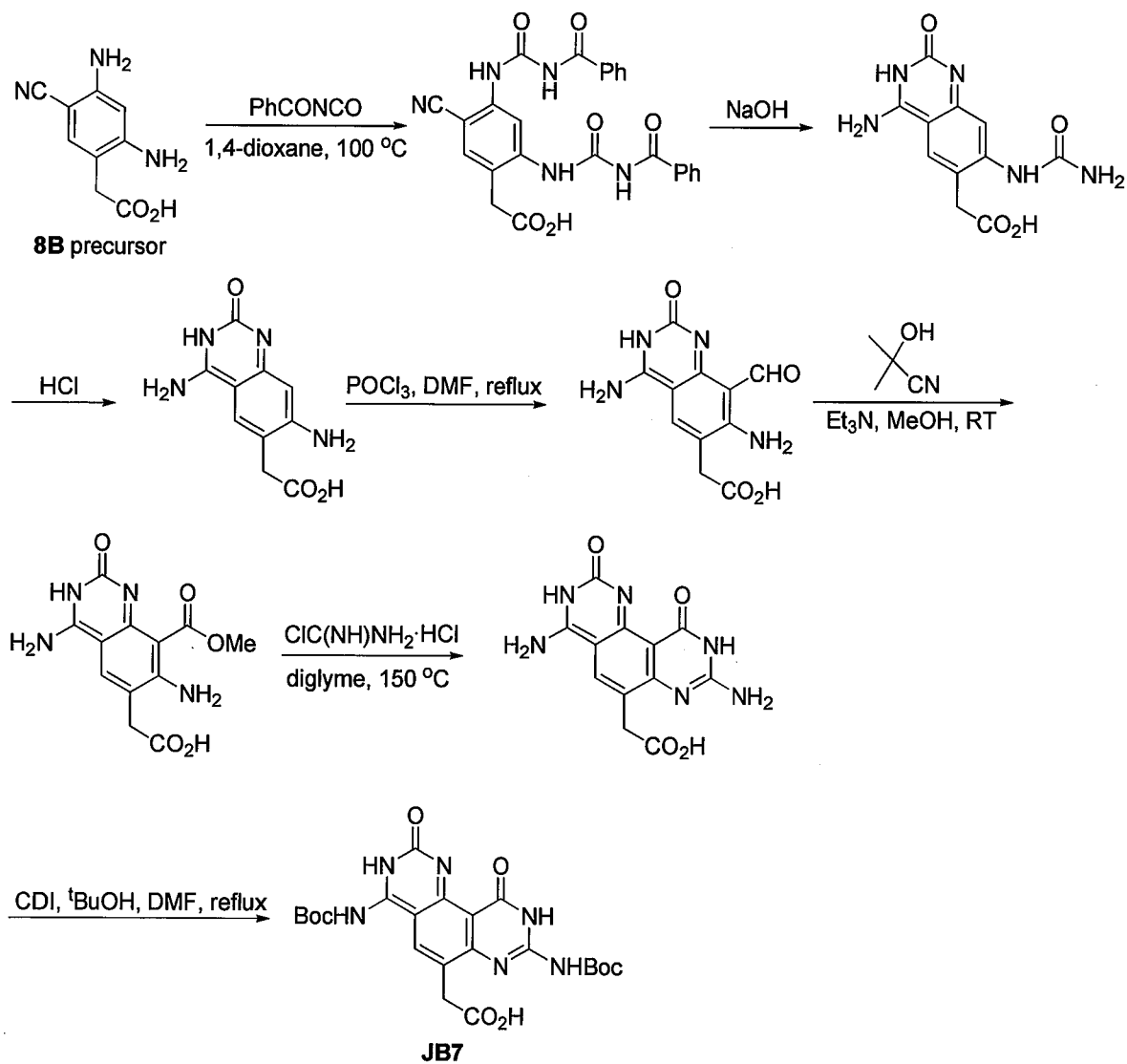


Fig. 7 (continued)

14/20

JB8

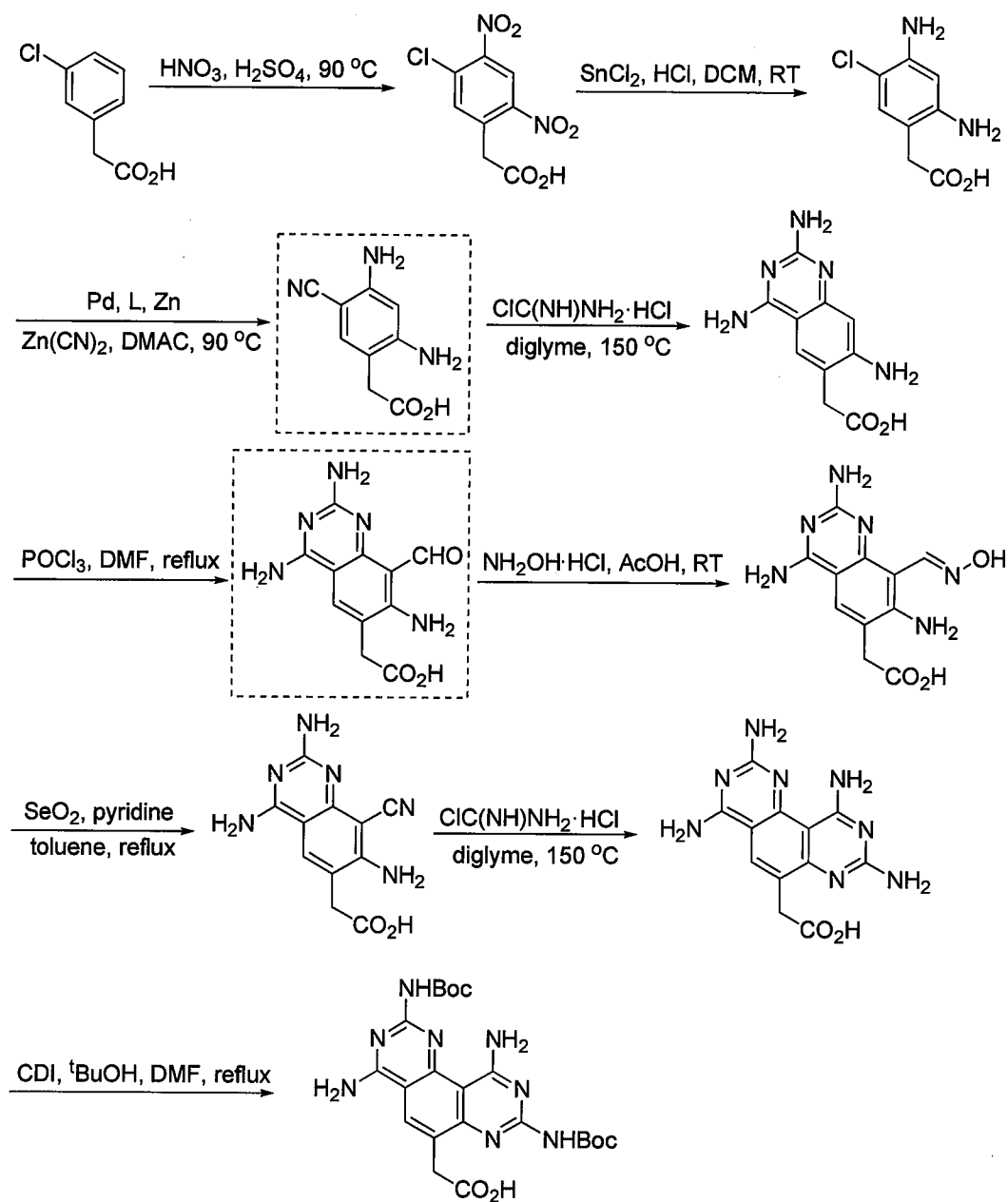
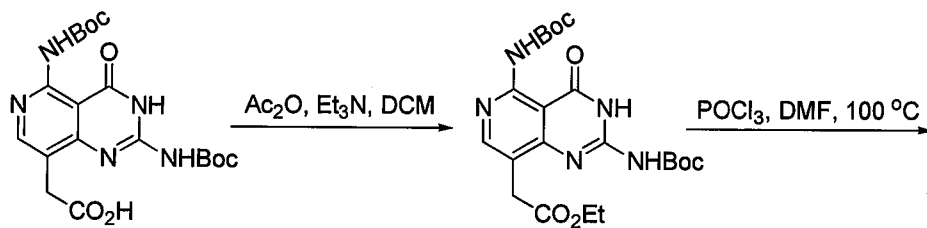


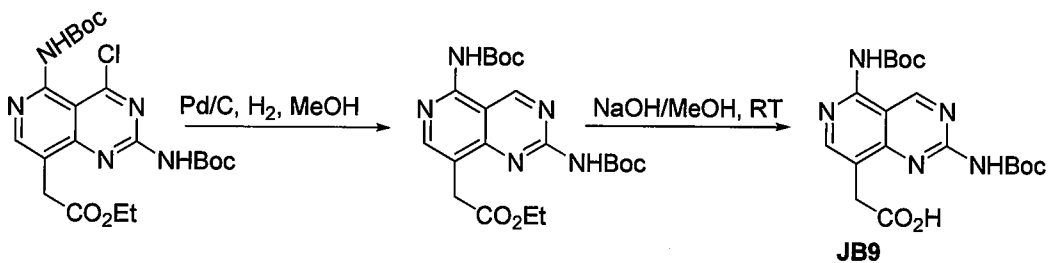
Fig. 7 (continued)

15/20

JB9



JB3



JB9

JB10

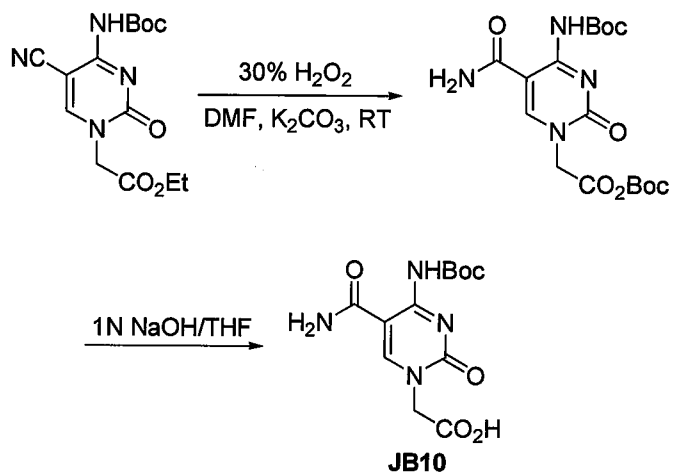


Fig. 7 (continued)

16/20

JB11

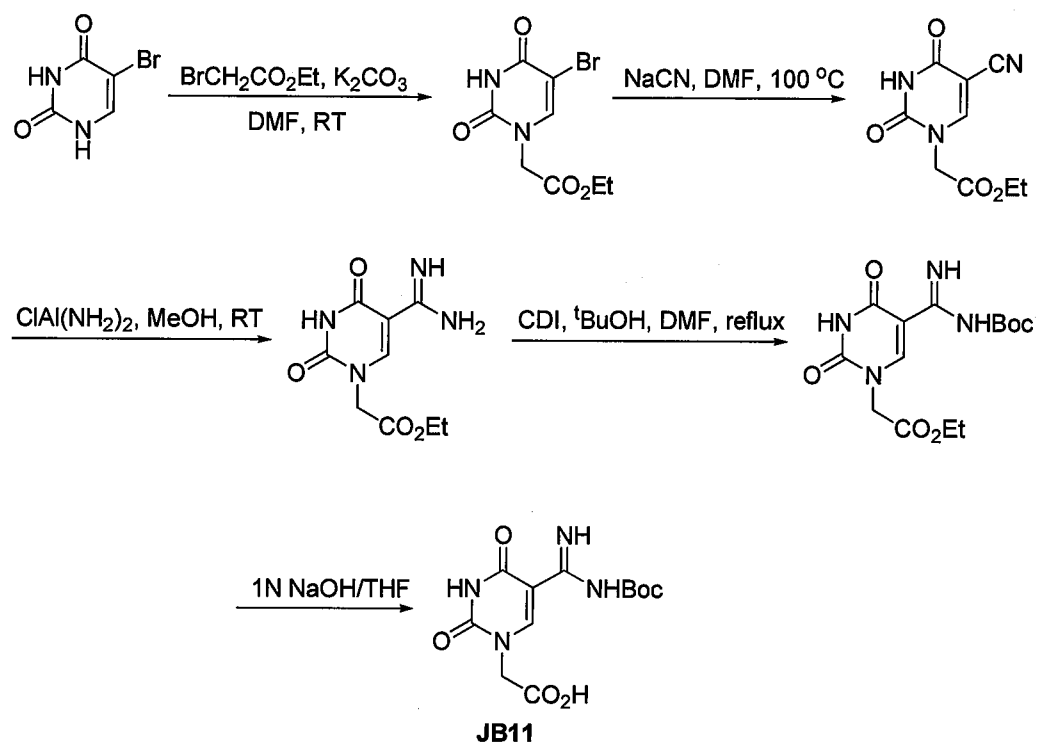
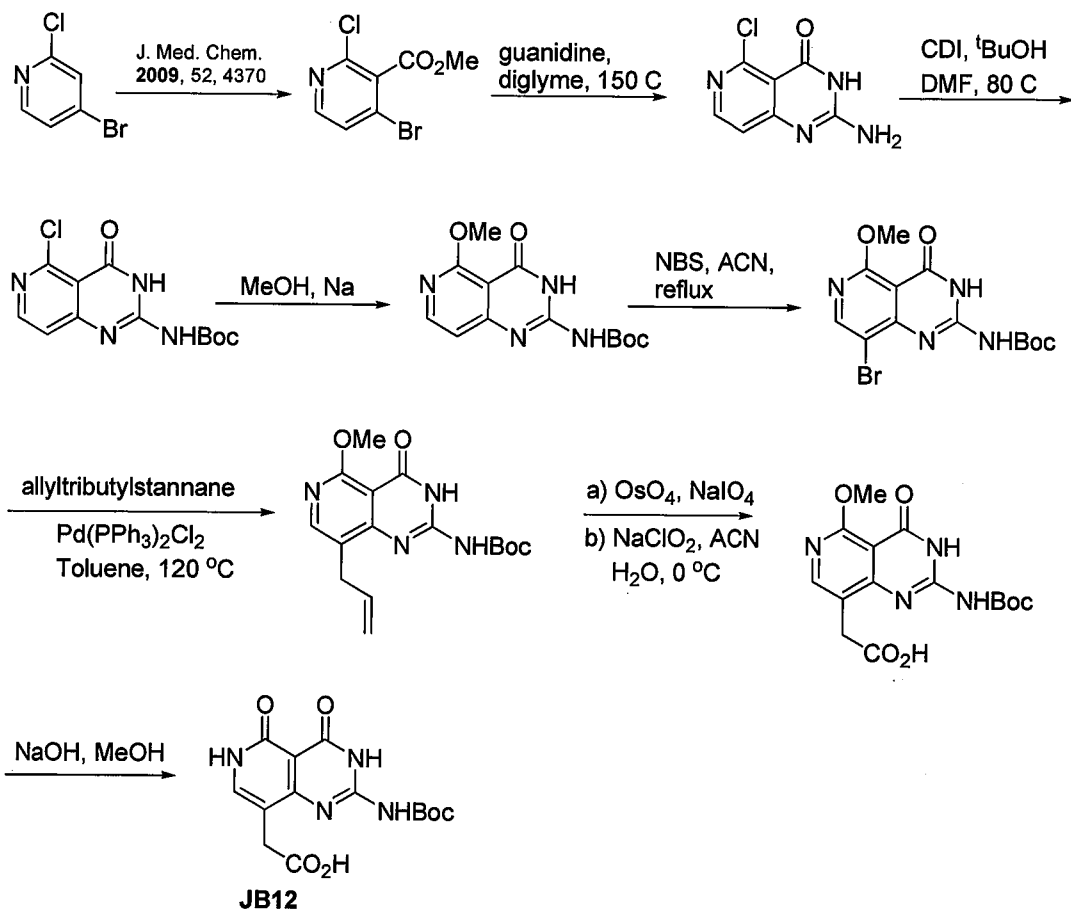


Fig. 7 (continued)

17/20

JB12



JB13

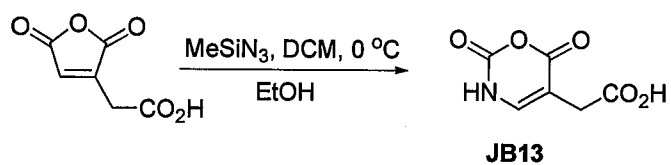
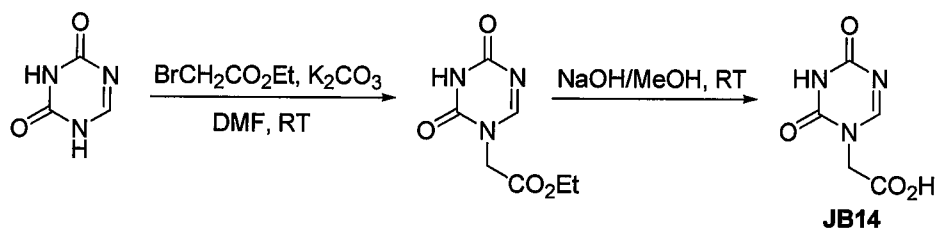


Fig. 7 (continued)

18/20

JB14



JB15

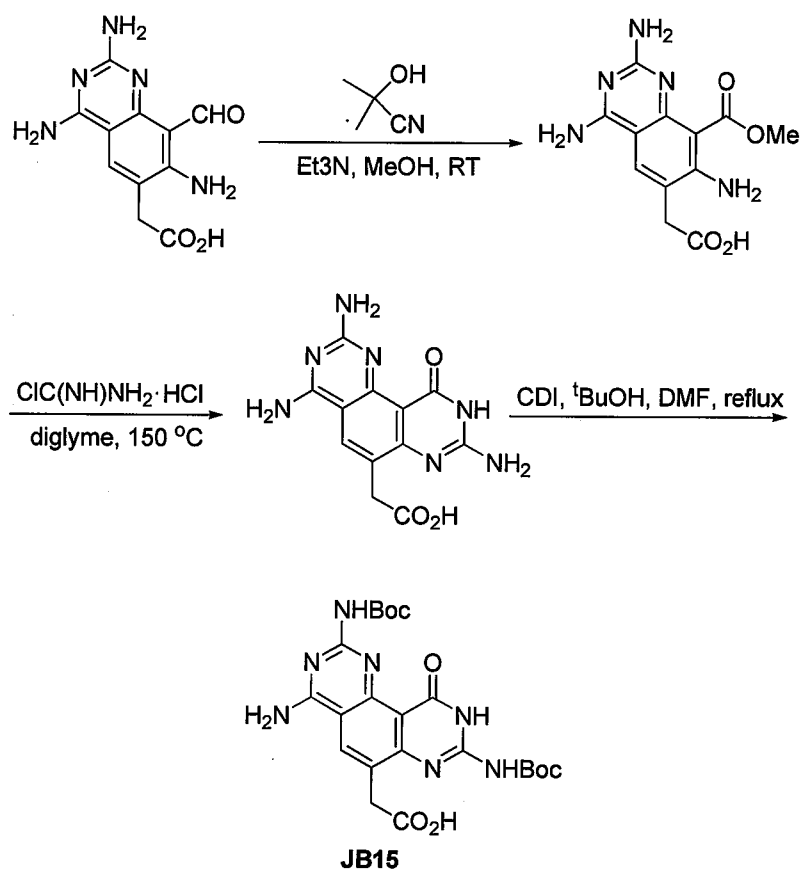


Fig. 7 (continued)

19/20

JB16

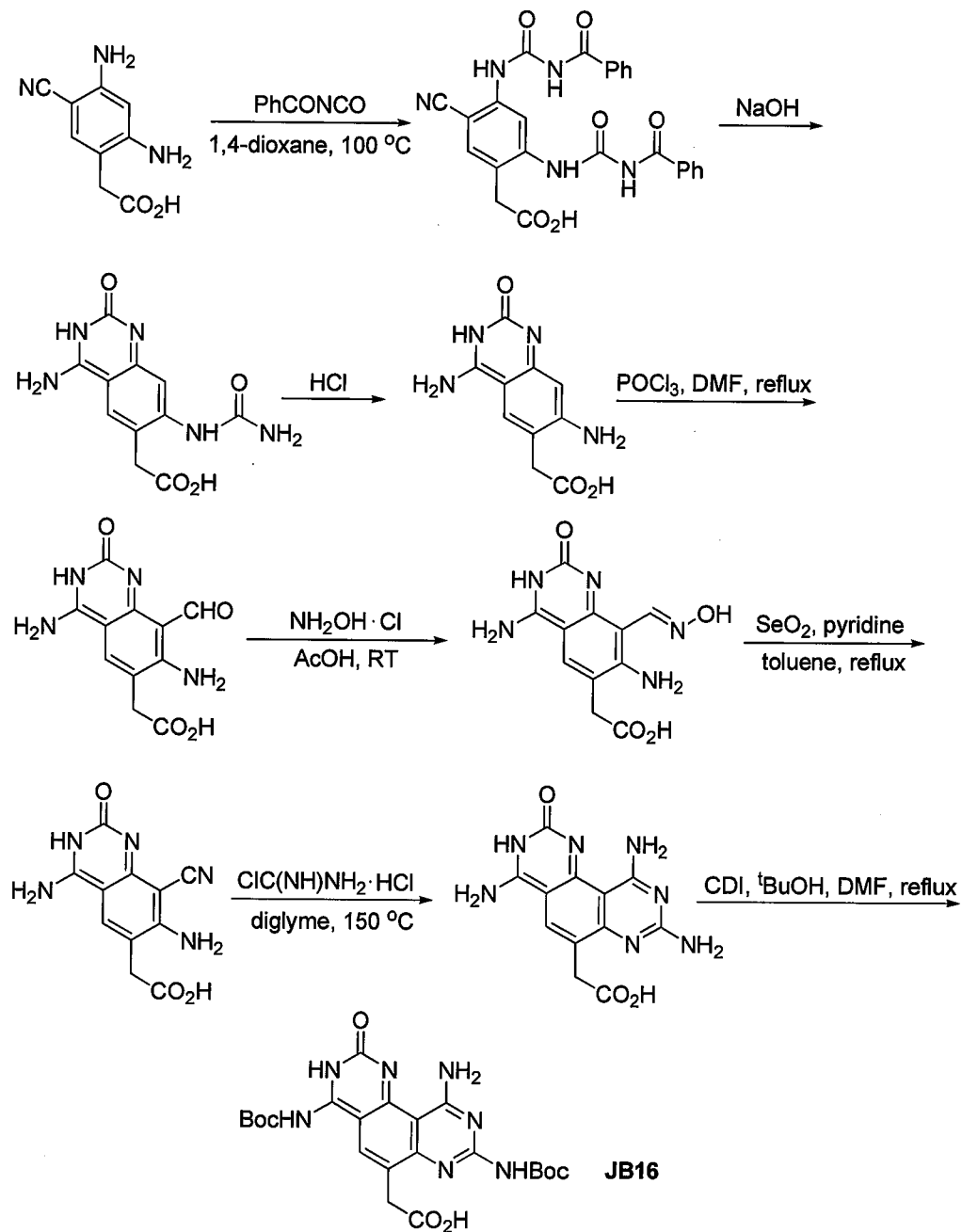
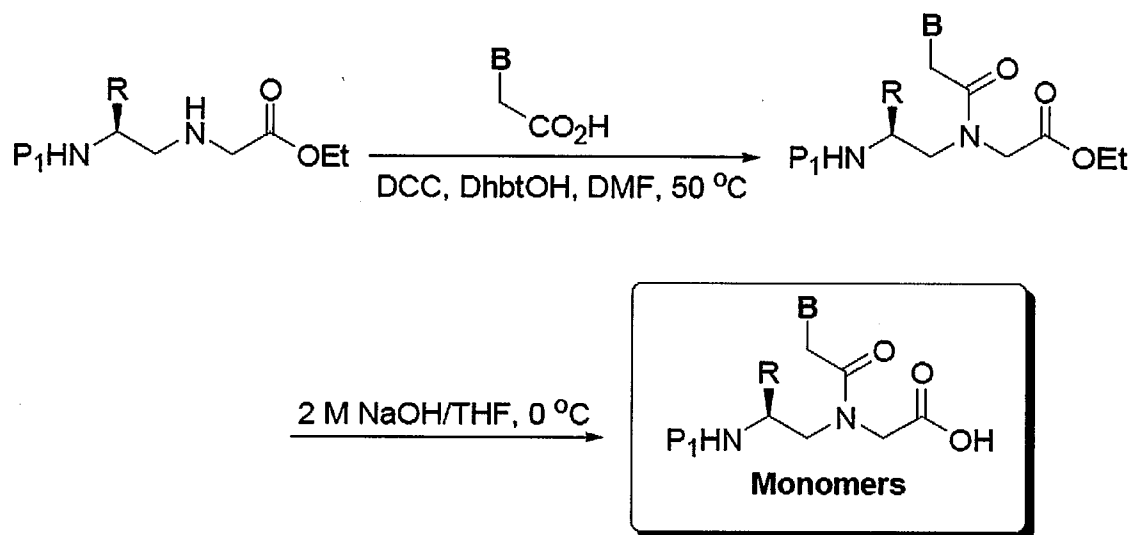
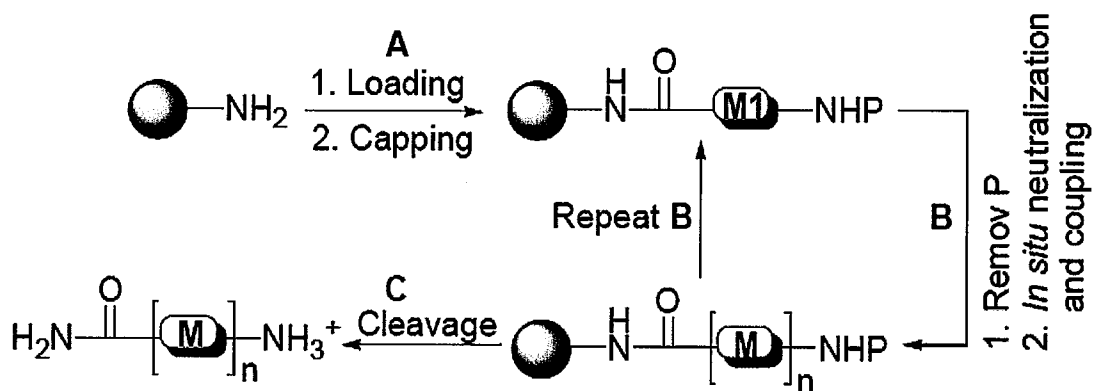


Fig. 7 (continued)

20/20

**Fig. 8****Fig. 9**