(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2007 (13.09.2007)

PCT

(10) International Publication Number
WO 2007/103307 A2

(51) International Patent Classification:
C12N 15/09 (2006.01)

(21) International Application Number:
PCT/US2007/005581

(22) International Filing Date: 5 March 2007 (05.03.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/779,375	3 March 2006 (03.03.2006)	US
60/779,376	3 March 2006 (03.03.2006)	US

(71) Applicant *(for all designated States except US)*: **CALIFORNIA INSTITUTE OF TECHNOLOGY** [US/US]; 1200 East California Boulevard, Pasadena, CA 91125 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **WANG, Pin** [CN/US]; 975 San Pasqual Street, Apt. #226, Pasadena, CA 91106 (US). **KWON, Inchan** [KR/US]; 600 Gooding Way #628, Albany, CA 94706 (US). **SON, Soojin** [KR/US]; 155 Washington Street #702, Jersey City, NJ 07302 (US). **TANG, Yi** [US/US]; 9141 Arcadia Avenue, San Gabriel, CA 91175 (US). **TIRRELL, David, A.** [US/US]; 714 Arden Road, Pasadena, CA 91106 (US).

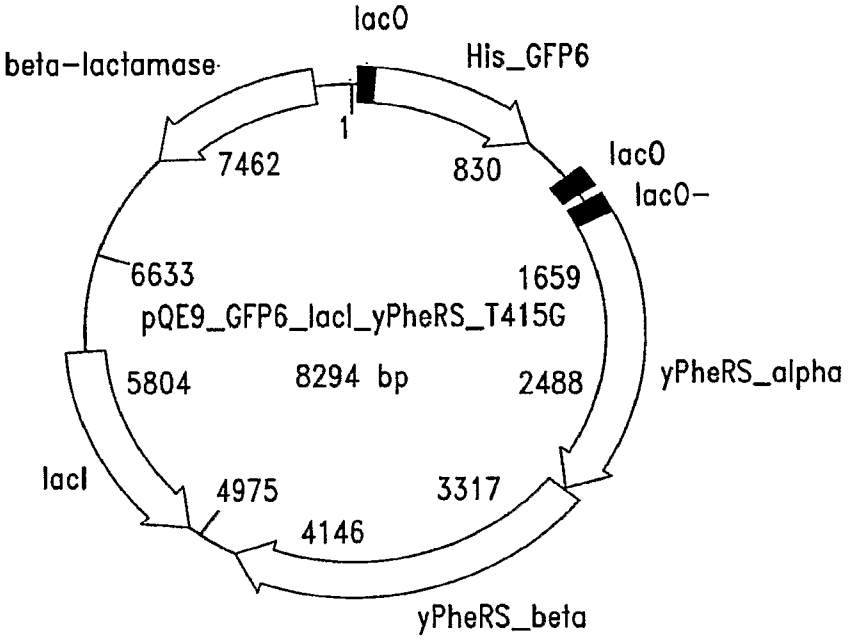
(74) Agents: **KITZAN HAINDFIELD, Melanie** et al.; Seed Intellectual PropertyLaw Group PLLC, 701 Fifth Avenue, Suite 5400, Seattle, WA 98104-7064 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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,

[Continued on next page]

(54) Title: SITE-SPECIFIC INCORPORATION OF AMINO ACIDS INTO MOLECULES



(57) Abstract: The invention provides certain embodiments relating to methods and compositions for incorporating non-natural amino acids into a polypeptid or protein by utilizing a mutant or modified aminoacyl-tRNA synthetase to charge the non-natural amino acid to a the corresponding tRNA. In certain embodiments, the tRNA is also modified such that the complex forms strict Watson-Crick base-pairing with a codon that normally forms wobble base- pairing with unmodified tRNA/aminoacyl-tRNA synthetase pairs.

WO 2007/103307 A2

GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SITE-SPECIFIC INCORPORATION OF AMINO ACIDS INTO MOLECULES

CROSS-REFERENCE TO RELATED APPLICATION

This application claims the benefit of the filing date of U.S. Provisional Application 60/779,375, filed on March 3, 2006, and U.S. Provisional Application 60/779,376, filed on March 3, 2006, the entire content of which are incorporated herein by reference.

STATEMENT OF GOVERNMENT INTEREST

This invention was made with federal government support under grant number GM62523, awarded by the NIH. The United States government has certain rights in the invention.

BACKGROUND OF THE INVENTION

Protein engineering is a powerful tool for modification of the structural catalytic and binding properties of natural proteins and for the *de novo* design of artificial proteins. Protein engineering relies on an efficient recognition mechanism for incorporating mutant amino acids in the desired protein sequences. Though this process has been very useful for designing new macromolecules with precise control of composition and architecture, a major limitation is that the mutagenesis is restricted to the 20 naturally occurring amino acids. However, it is becoming increasingly clear that incorporation of unnatural amino acids can extend the scope and impact of protein engineering methods.

Non-natural amino acids carrying a wide variety of novel functional groups have been globally replaced for residue-specific replacement or incorporation into recombinant proteins. Biosynthetic assimilation of non-canonical amino acids into proteins has been achieved largely by exploiting the capacity of the wild type synthesis apparatus to utilize analogs of naturally occurring amino acids (Budisa 1995, *Eur. J. Biochem* 230: 788-796; Deming

1997, *J. Macromol. Sci. Pure Appl. Chem* A34; 2143-2150; Duewel 1997, *Biochemistry* 36: 3404-3416; van Hest and Tirrell 1998, *FEBS Lett* 428(1-2): 68-70; Sharma *et al.*, 2000, *FEBS Lett* 467(1): 37-40). However, there are situations in which single-site substitution or incorporation by non-natural amino acids is required. Such a methodology would enable the tailoring in a protein (the size, acidity, nucleophilicity, hydrogen-bonding or hydrophobic properties, etc. of amino acids) to fulfill a specific structural or functional property of interest. The ability to site-specifically incorporate such amino acid analogs into proteins would greatly expand our ability to rationally and systematically manipulate the structures of proteins, both to probe protein function and create proteins with new properties. For example, the ability to synthesize large quantities of proteins containing heavy atoms would facilitate protein structure determination, and the ability to site specifically substitute fluorophores or photo-cleavable groups into proteins in living cells would provide powerful tools for studying protein functions *in vivo*.

In recent years, several laboratories have pursued an expansion in the number of genetically encoded amino acids, by using either a nonsense suppressor or a frame-shift suppressor tRNA to incorporate non-canonical amino acids into proteins in response to amber or four-base codons, respectively (Bain *et al.*, *J. Am. Chem. Soc.* 111: 8013, 1989; Noren *et al.*, *Science* 244: 182, 1989; Furter, *Protein Sci.* 7: 419, 1998; Wang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, 100: 56, 2003; Hohsaka *et al.*, *FEBS Lett.* 344: 171: 1994; Kowal and Oliver, *Nucleic Acids Res.* 25: 4685, 1997). Such methods insert non-canonical amino acids at codon positions that will normally terminate wild-type peptide synthesis (*e.g.*, a stop codon or a frame-shift mutation). These methods have worked well for *single-site* insertion of novel amino acids. However, their utility in *multisite* position specific (versus residue specific) substitution or incorporation is limited by modest (20-60%) suppression efficiencies (Anderson *et al.*, *J. Am. Chem. Soc.* 124: 9674, 2002; Bain *et al.*, *Nature* 356: 537, 1992; Hohsaka *et al.*, *Nucleic Acids Res.* 29: 3646, 2001).

This is so partially because too high a stop codon suppression efficiency will interfere with the normal translation termination of some non-targeted proteins in the organism. On the other hand, a low suppression efficiency will likely be insufficient to suppress more than one nonsense or frame-shift mutation sites in the target protein, such that it becomes more and more difficult or impractical to synthesize a full-length target protein incorporating more and more non-canonical amino acids.

Efficient multisite incorporation has been accomplished by replacement of natural amino acids in auxotrophic *Escherichia coli* strains, for example, by using aminoacyl-tRNA synthetases with relaxed substrate specificity or altered editing activity (Wilson and Hatfield, *Biochim. Biophys. Acta* 781: 205, 1984; Kast and Hennecke, *J. Mol. Biol.* 222: 99, 1991; Ibba *et al.*, *Biochemistry* 33: 7107, 1994; Sharma *et al.*, *FEBS Lett.* 467: 37, 2000; Tang and Tirrell, *Biochemistry* 41: 10635, 2002; Datta *et al.*, *J. Am. Chem. Soc.* 124: 5652, 2002; Doring *et al.*, *Science* 292: 501, 2001). Although this method provides efficient incorporation of analogues at multiple sites, it suffers from the limitation that the novel amino acid must "share" codons with one of the natural amino acids. Thus for any given codon position where both natural and novel amino acids can be inserted, other than a probability of incorporation, there is relatively little control over which amino acid will end up being inserted. This may be undesirable, since for an engineered enzyme or protein, non-canonical amino acid incorporation at an unintended site may unexpectedly compromise the function of the protein, while missing incorporating the non-canonical amino acid at the designed site will fail to achieve the design goal.

In general, multisite substitution methods are relatively simple to carry out, but all sites corresponding to a particular natural amino acid throughout the protein are replaced. The extent of incorporation of the natural and non-natural amino acid may also vary. Furthermore, multisite incorporation of analogs often results in toxicity when cells are utilized, which makes it difficult

to study the mutant protein in living cells. The present invention overcomes these hurdles by allowing for site-specific mutation of amino acids in proteins.

Certain embodiments disclosed herein provide a new technique for the incorporation of replacement amino acids, including naturally occurring amino acids, or non-standard or non-canonical amino acids into proteins that is based on breaking the degeneracy of the genetic code. Specifically, certain embodiments herein allow for high fidelity position-specific substitution or incorporation of non-natural amino acids into proteins.

BRIEF SUMMARY OF THE INVENTION

Certain embodiments disclosed herein provide for compositions of components used in protein biosynthetic machinery, which include external mutant aminoacyl tRNA molecules, external mutant aminoacyl-tRNA synthetase (AARS) molecules, or pairs of the same, as well as the individual components of the pairs. As disclosed herein inter alia, external mutant molecules are

Methods are also provided for generating and selecting external mutant tRNAs, external mutant aminoacyl-tRNA synthetases, and pairs thereof that are capable of incorporating amino acids, including non-natural amino acids, into polypeptides or proteins. Certain compositions of specific embodiments include novel external mutant tRNA or external mutant aminoacyl-tRNA synthetase pairs. The novel external mutant tRNA molecules, AARS molecules, or AARS-tRNA pairs can be used to incorporate an unnatural amino acid in a polypeptide *in vitro* and *in vivo*. Other embodiments of the invention include selecting external mutant pairs.

Some compositions of the present invention include an external mutant aminoacyl-tRNA synthetase, where the external mutant tRNA synthetase preferentially aminoacylates an external mutant tRNA with an unnatural amino acid, optionally, *in vivo*. In one embodiment, a nucleic acid or

polynucleotide encoding an external mutant synthetase is provided, or a complementary nucleic acid sequence thereof.

Thus, certain embodiments include a composition comprising a first vector containing a polynucleotide encoding a modified aminoacyl tRNA synthetase (AARS), wherein said polynucleotide modified synthetase is mutated at one or more codons encoding the amino acid binding region necessary for interaction with the amino acid to be paired with a tRNA molecule, and wherein said modified synthetase is capable of charging a tRNA molecule with a non-natural amino acid. In some embodiments, the binding region comprises no more than 30, 20, 15, 10, or 5 contiguous amino acid residues. In at least one embodiment, the modified AARS is selected from the group consisting of a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS. In some embodiments wherein the modified AARS is a modified PheRS, said PheRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 412, 415, 418, and 437. In at least one embodiment wherein said modified AARS is a modified TrpRS, the TrpRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 4, 5, 7, 132, 133, 141, and 143. In some embodiments wherein the modified AARS is a modified MetRS, the MetRS is mutated at amino acid sequence position number 13.

At least one embodiment further comprises a second vector containing a polynucleotide encoding a tRNA molecule. In at least one embodiment, said first and second vectors are the same vector. In other embodiments, said first and second vectors are different vectors.

In at least one embodiment, the tRNA is endogenous, and in at least one embodiment, the tRNA is modified. In at least one embodiment, the tRNA is modified such that it contains a mutated anticodon that base pairs with a corresponding wobble degenerate codon with an affinity greater than the affinity of the natural tRNA. In some embodiments, the AARS and the tRNA are

from the same or different organisms. In at least one embodiment, the non-natural amino acid is selected from the group consisting of: azidonorleucine, 3-(1-naphthyl)alanine, 3-(2-naphthyl)alanine, *p*-ethynyl-phenylalanine, *p*-propargly-oxy-phenylalanine, *m*-ethynyl-phenylalanine, 6-ethynyl-tryptophan, 5-ethynyl-troptophan, (R)-2-amino-3-(4-ethynyl-1H-pyrol-3-yl)propanic acid, *p*-bromophenylalanine, *p*-idiophenylalanine, p-azidophenylalanine, 3-(6-chloroindolyl)alanine, 3-(6-bromoindolyl)alanine, 3-(5-bromoindolyl)alanine, azidohomoalanine, and *p*-chlorophenylalanine.

Other embodiments disclosed herein include a polypeptide comprising a modified aminoacyl tRNA synthetase (AARS), wherein said modified synthetase is mutated at one or more codons in the amino acid binding region necessary for interaction with the amino acid to be paired with a tRNA molecule, and wherein said modified synthetase is capable of charging a tRNA molecule with a non-natural amino acid. In at least one embodiment, the binding region comprises no more than 30, 20, 15, 10, or 5 contiguous amino acid residues.

In at least one embodiment, the modified AARS is selected from the group consisting of a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS. In some embodiments wherein the modified AARS is a modified PheRS, said PheRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 412, 415, 418, and 437. In at least one embodiment wherein said modified AARS is a modified TrpRS, the TrpRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 4, 5, 7, 132, 133, 141, and 143. In some embodiments wherein the modified AARS is a modified MetRS, the MetRS is mutated at amino acid sequence position number 13.

Certain embodiments include translation system comprising the polynucleotide encoding a modified aminoacyl tRNA synthetase (AARS), wherein said polynucleotide modified synthetase is mutated at one or more

codons encoding the amino acid binding region necessary for interaction with the amino acid to be paired with a tRNA molecule, and wherein said modified synthetase is capable of charging a tRNA molecule with a non-natural amino acid. In at least one embodiment, the system comprises a host cell. In at least one embodiment, the modified aminoacyl tRNA synthetase is derived from an organism different than the host cell. In another embodiment, the translation system further comprises a polynucleotide encoding a modified tRNA molecule.

In certain embodiments, the modified tRNA molecule is derived from an organism different than the host cell. In certain embodiments, the modified tRNA molecule is derived from a eukaryotic cell and the host cell is a prokaryotic cell. In still other embodiments, the cell is an auxotroph.

In some embodiments, the translation system further comprises a culture media containing one or more non-natural amino acids. In still other embodiments, said one or more non-natural amino acids are selected from the group consisting of: azidonorleucine, 3-(1-naphthyl)alanine, 3-(2-naphthyl)alanine, *p*-ethynyl-phenylalanine, *p*-propargly-oxy-phenylalanine, *m*-ethynyl-phenylalanine, 6-ethynyl-tryptophan, 5-ethynyl-troptophan, (R)-2-amino-3-(4-ethynyl-1H-pyrol-3-yl)propanic acid, *p*-bromophenylalanine, *p*-idiophenylalanine, p-azidophenylalanine, 3-(6-chloroindolyl)alanine, 3-(6-bromoindolyl)alanine, 3-(5-bromoindolyl)alanine, azidohomoalanine, and *p*-chlorophenylalanine. In still other embodiments, said modified AARS is selected from the group consisting of: a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS.

Other embodiments relate to a method for incorporating a non-natural amino acid into a target polypeptide at one or more specified position(s), the method comprising the steps of:

(1) determining the structural change in the polypeptide for incorporation of a non-natural at one specific position in the polypeptide;

(2) providing a translation system;

(3) providing to the translation system a first polynucleotide of claim 1, or the modified AARS encoded thereby;

(4) providing to the translation system the non-natural amino acid;

(5) providing to the translation system a template polynucleotide encoding a polypeptide of interest, and,

(6) allowing translation of the template polynucleotide, thereby incorporating the non-natural amino acid into the polypeptide of interest at the specified position(s),

wherein steps (1)-(4) are effectuated in any order.

In certain embodiments, said translation system comprises a cell. In some embodiments, step (4) is effectuated by contacting said translation system with a solution containing the non-natural amino acid. In at least one embodiment, the specificity constant (k_{cat} / K_M) for activation of said non-natural amino acid by said modified AARS is at least 5-fold larger than that for said natural amino acid. In certain embodiments, the modified AARS mischarges a tRNA at a rate of no more than 1%, 2%, 3%, 4%, 5%, 6%, 7%, or 8%. In still other embodiments, the tRNA is a modified tRNA. In certain embodiments, said first polynucleotide or said second polynucleotide further comprises either a constitutively active or an inducible promoter sequence that controls the expression of the tRNA or AARS. In at least one embodiment, the method further comprises the step of screening for cells containing a modified AARS. In another embodiment, the method further comprises the step of verifying the incorporation of the non-natural amino acid. In another embodiment, the modified AARS is selected from the group consisting of: PheRS, TyrRS, TrpRS, and MetRS. Still other embodiments comprise a polypeptide made by the method disclosed.

Certain embodiments disclosed herein include a method for incorporating at least one non-natural amino acid into a target polypeptide at one or more specified location(s), the method comprising providing a translation system containing at least one non-natural amino acid; providing to the

translation system one or more modified AARS selected from the group consisting of: modified PheRS, TrpRS, TyrRS, and MetRS; providing to the translation system a polynucleotide encoding a target polypeptide of interest; and allowing translation of interest, thereby incorporating at least one non-natural amino acid into the target polypeptide. Certain embodiments include a polypeptide made by this method.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

Figure 1 shows a sequence alignment of PheRS variants from conserved sequences of *Thermus thermophilus*, *Escherichia coli*, and *Saccharomyces cerevisiae*.

Figure 2 shows several exemplary amino acids (naturally occurring or non-natural) used for some embodiments disclosed herein.

Figure 3 shows the amino acid sequence for an exemplary polypeptide used for some embodiments disclosed herein, dihydrofolate reductase (DHFR). Four proteolytic peptide fragments (labeled Peptide A, Peptide B, Peptide C and Peptide D) were used for MALDI and liquid chromatography-Mass Spectrum/Mass Spectrum (LC-MS/MS) analyses as underscored.

Figure 4 shows a MALDI-MS of proteolytic peptide fragments derived from mDHFR expressed in media supplemented with (a) amino acid 1 (3 mM); (b) amino acid 7 (3 mM) and 1 (0.03 mM); (c) amino acid 2 (3 mM) and 1 (0.03); (d) amino acid 2 (3 mM) and 1 (0.03mM). No tryptophan is supplemented during induction, except that 1mM of tryptophan is supplemented in media at (c). Peptide B, containing one Phe codon, is the control. Peptide A contains an amber codon (Z), the amino acid for which is assigned based on the mass units for Peptide A at different expression conditions. Due to lysine incorporation with Peptide A (c), C-terminal lysine was cleaved to produce a shorter Peptide A (NGDLPWPPLRNEK) (SEQ ID NO: 4).

Figure 5 Tandem mass spectrum of Peptide A (NGDLPWPPLRNEZK) (SEQ ID NO: 5) derived from DHFR expressed in

media supplemented with amino acid 7 (3mM) and amino acid 1 (0.03 mM). Partial sequence of PWPPLRNE (SEQ ID NO: 6) and residue Z (corresponding to amino acid 7) of Peptide A can be assigned from the annotated y and b ion series, respectively.

Figure 6 shows MALDI-MS of proteolytic peptide fragments derived from mDHFR expressed in media supplemented with (a) amino acid 2 (3 mM) and amino acid 6 (3 mM) and amino acid 1 (0.03 mM) and amino acid 13 (0.2 mM); (b) amino acid 2 (3 mM) and amino acid 6 (0.01 mM) and amino acid 1 (0.03 mM) and amino acid 13 (0.2 mM); (c) amino acid 3 (3 mM) and amino acid 6 (0.01 mM) and amino acid 1 (0.03 mM) and amino acid 13 (0.01 mM). Peptide C, with one Phe codon, is the control.

Figure 7 shows MALDI-MS results of proteolytic peptide fragments derived from mDHFR expressed in media supplemented with (a) amino acid 9 (3 mM) and amino acid 6 (0.03 mM) and amino acid 1 (0.01 mM); (b) amino acid 10 (3 mM) and amino acid 6 (0.03 mM) and amino acid 1 (0.03 mM); (c) amino acid 11 (3 mM) and amino acid 6 (0.01 mM) and amino acid 1 (0.03 mM). Peptide D was the control.

Figure 8 shows the aminoacylation of yeast $tRNA^{Phe}_{CUA}$ (square) and $tRNA^{Phe}_{CUA\_UG}$ (circle) with lysine by eLysS.

Figure 9 shows LC-MS chromatograms of tryptic digests of the mDHFR polypeptides

Figure 10 shows ATP-PPi exchange rates for phenylalanine, tryptophan and p-bromophenylalanine by wild type yeast PheRS and external mutant yeast PheRS T415G or external mutant yeast PheRS T415A.

Figure 11 shows aminoacylation of phenylalanine and tryptophan by wild type aminoacyl tRNA synthetase or external mutant tRNA synthetase (T415G or T415A).

Figure 12 shows incorporation of lysine (open square), tryptophan (cross-hatch) or p-bromophenylalanine (checker board).

Figure 13 shows mass spectra of p-ethynylphenylalanine incorporated into a polypeptide using a modified tRNA synthetase (T415G) and $tRNA^{Phe}$ with an amber suppressor in a host cell.

Figure 14 shows a FACS of green fluorescent protein (GFP) incorporation of amino acids using amber suppression codon in a test protein.

Figure 15 illustrates an exemplary plasmid mapping of a mutant aminoacyl tRNA synthetase (T415G).

Figure 16 illustrates exemplary mutations made in a yeast phenylalanine tRNA synthetase.

DETAILED DESCRIPTION OF THE INVENTION

Proteins are at the crossroads of virtually every biological process, from photosynthesis and vision to signal transduction and the immune response. Modifying proteins or polypeptides to include non-natural amino acids has great potential for use in human therapeutics, agriculture, biofuel, and other areas.

Aminoacyl-tRNA synthetases catalyze the aminoacylation reaction for incorporation of amino acids into proteins via the corresponding transfer RNA molecules. Precise manipulation of synthetase activity can alter the aminoacylation specificity to stably attach non-canonical amino acids into the intended tRNA. Then, through codon-anticodon interaction between message RNA (mRNA) and tRNA, the amino acid analogs can be delivered into a growing polypeptide chain. Thus, incorporation of non-natural amino acids into proteins relies on the manipulation of amino acid specificity of aminoacyl tRNA synthetases (AARS).

Aminoacyl-tRNA synthetases function to transform the genetic code sequences into biologically functional proteins through a two-step aminoacylation reaction. As an initial step, the cognate amino acid is activated by AARS in the presence of ATP to form the amino acid adenylate; subsequently AARS catalyzes the esterification reaction to join the amino acid

to 2'- or 3'-OH of the terminal ribonucleotide of its cognate tRNA. Once the aminoacylation reaction occurs, the amino acid is directed into the growing polypeptide chain by the charged tRNA.

Certain embodiments disclosed herein relate to mutant or modified aminoacyl tRNA synthetase (AARS or RS) molecules that have been mutated or modified such that the enzymes are capable of charging a tRNA molecule with a replacement amino acid, preferably a non-natural amino acid due to disruption of between the synthetase and the corresponding natural amino acid.

For example, the disruption may be due to interfering with Watson-Crick base pairing, interfering with wobble base pairing, or creation of novel wobble or other base pairing.

Some embodiments relate to a polynucleotide encoding a mutant or modified tRNA of a tRNA for a natural amino acid, wherein the natural amino acid is encoded by one or more wobble degenerate codon(s), the modified tRNA comprises a modified anticodon sequence that forms Watson-Crick base-pairing with one of the wobble degenerate codon(s). Preferably, the modified tRNA is not or only inefficiently charged by an endogenous aminoacyl-tRNA synthetase (AARS) for the natural amino acid. In one embodiment, multiple modified AARS molecules may be used with one or more tRNA molecules. In one embodiment, one or more modified or mutated AARS molecule can be used with one or more native tRNA molecule, while in another embodiment a modified or mutated AARS can be used with a modified or mutated tRNA molecule. In certain embodiments, one or more pairs of modified AARS/tRNA molecules may be utilized. In certain embodiments, heterologous pairs may be used. In certain embodiments, one or more modified or mutant AARS and/or tRNA may be derived from the same or a different organism.

In some exemplary embodiments, a particular AARS may utilize several methods for incorporation of a replacement amino acid (including a non-natural amino acid) into a polypeptide or protein. For example, a single

AARS may utilize a nonsense codon (such as an amber stop codon) for incorporation of a replacement amino acid (such as a non-natural amino acid) at a particular location in the polypeptide. In addition or instead of this, a wobble codon (such as UUU) could be used for incorporation of the replacement amino acid at the wobble codon site (in this example, using a modified PheRS).

In other exemplary embodiments, multiple replacement amino acids (such as two different non-natural amino acids) may be incorporated into a polypeptide or protein through the use of various methods. For example, one non-natural amino acid may be incorporated at a wobble site, while a different non-natural amino acid may be incorporated at an amber stop codon. In some exemplary embodiments, incorporation of multiple replacement amino acids (including non-natural amino acids) includes utilizing one AARS for multiple different amino acid analogs of any amino acid, or multiple different amino acid analogs all of a particular naturally occurring amino acid. Thus, for example, a modified PheRS may be used to incorporate multiple different phenylalanine analogs, such as bromophenylalanine and/or p-idiophenylalanine, in the same polypeptide or protein.

A similar approach involves using a heterologous synthetase and a mutant initiator tRNA of the same organism or a related organism as a tRNA molecule. (See, for example, Kowal, *et al.*, *PNAS*, 98, 2268 (2001)).

In certain embodiments, the modified or mutated RS interacts with the desired amino acid replacement (whether naturally occurring or non-natural amino acid) with an altered binding specificity and/or altered catalytic event of the enzyme toward the amino acid replacement when compared to the wild type RS enzyme or wild type corresponding amino acid.

In enzyme kinetics, k_{cat} is a first-order rate constant corresponding to the slowest step or steps in the overall catalytic pathway. The k_{cat} represents the maximum number of molecules of substrate which can be converted into product per enzyme molecule per unit time (which occurs if the enzyme is

"saturated" with substrate), and thus is often referred to as the turnover number. The K_m is an apparent dissociation constant and is related to the enzyme's affinity for the substrate; it is the product of all the dissociation and equilibrium constants prior to the first irreversible step in the pathway. Often, it is a close measure of the enzyme-substrate dissociation constant. The k_{cat}/K_m is a second-order rate constant which refers to the free enzyme (not enzyme-substrate complex) and is also a measure of the overall efficiency of the enzyme catalysis and is also referred to as the specificity constant.

In certain embodiments, the external mutant synthetase has improved or enhanced enzymatic properties, *e.g.*, the K_m is higher or lower, the k_{cat} is higher or lower, the value of k_{cat}/K_m is higher or lower or the like, for the unnatural amino acid compared to a naturally occurring amino acid, e.g., one of the 20 known amino acids. The Km of the mutant or modified AARS is preferably equal or lower for the non-natural amino acid than for the corresponding wild type natural amino acid.

In certain embodiments, the k_{cat}/K_m values of the RS variant may range from 3-fold, 5-fold, 10-fold, 25-fold, 50-fold, 100-fold, 150-fold, 200-fold, 250-fold, 300-fold, 350-fold, 385-fold, 400-fold higher than for the naturally occurring amino acid.

In certain embodiments, the modified tRNA interacts with the wobble degenerate codon with an affinity at 37°C of at least about 1.0 kcal/mole, 1.5 kcal/mole, 2.0 kcal/mole, 2.5 kcal/mole, 3.0 kcal/mole, 3.5 kcal/mole, 4.0 kcal/mole, 4.5 kcal/mole, 5.0 kcal/mole or greater (or any value therebetween) favorably than the interaction between its unmodified version and the wobble degenerate codon.

For example, phenylalanine (Phe) is encoded by two codons, UUC and UUU. Both codons are read by a single tRNA, which is equipped with the anticodon sequence GAA. The UUC codon is therefore recognized through standard Watson-Crick base-pairing between codon and anticodon; UUU is read through a G-U wobble base-pair at the first position of the anticodon

(Crick, *J. Mol. Biol.* 19: 548, 1966; Soll and RajBhandary, *J. Mol. Biol.* 29: 113, 1967). Thermal denaturation of RNA duplexes has yielded estimates of the Gibbs free energies of melting of G-U, G-C, A-U, and A-C basepairs as 4.1, 6.5, 6.3, and 2.6 kcal/mol, respectively, at 37°C. Thus the wobble basepair, G-U, is less stable than the Watson-Crick basepair, A-U. A modified $tRNA^{Phe}$ outfitted with the AAA anticodon ($tRNA^{Phe}_{AAA}$) was engineered to read the UUU codon, and was predicted to read such codons faster than wild-type $tRNA^{Phe}_{GAA}$.

In some embodiments, the binding pocket of the RS is modified such that the modified RS exhibits a preference for the non-natural amino acid over the corresponding naturally occurring amino acid. In preferred embodiments, the RS is modified at one or more codon necessary for structural contact between the RS and the amino acid being charged to the tRNA. In certain embodiments, the one or more codon selected for mutation or modification are selected by way of computer modeling. While any RS can be modified according to the present disclosure, certain embodiments relate to phenylalaanyl-tRNA synthetase (PheRS), or tryptophan tRNA synthetase (TrpRS). In some embodiments, the modified RS is from *Saccharomyces cerevisiae*, or another eukaryotic cell. In other embodiments, the modified RS is from *E. coli* or another prokaryotic cell. In certain embodiments wherein the RS is a PheRS, the enzyme has a point mutation (N412G), (T415G), (T415A), (S418C), or (S437F) in the alpha subunit of the enzyme, or mutations at equivalent locations or positions in a homologous protein of another species or organism. The point mutations (for example, the T to G or A mutation at position 415, or S to C mutation at position 418, or N to G mutation at position 412, or S to F mutation at position 437) are located in the binding pocket region of the aminoacyl-tRNA synthetase (RS).

In some exemplary embodiments, typical Km values for different analogs with AARS may range from approximately 15 microM, 20 microM, 30 microM, 50 microM, 75 microM, 100 microM, 150 microM, 200 microM, 300 microM, 400 microM, 440 microM, 500 microM, 1000 microM, 1500 microM,

2000 microM, 3000 microM, 4000 microM, 5000 microM, 6000 microM, or greater or any value therebetween.

Likewise, the k_{cat} values of the mutant AARS is preferably equal to or higher for the amino acid analog than for the natural amino acid. For example, k_{cat} values for different analogs with the corresponding AARS may range from approximately 0.002 sec^{-1}, 0.0018 sec^{-1}, 0.0015 sec^{-1}, 0.014 sec^{-1}, 0.1 sec^{-1}, 0.3 sec^{-1}, 1 sec^{-1}, 3 sec^{-1}, 5 sec^{-1}, 8 sec^{-1}, 10 sec^{-1}, 13.3 sec^{-1}, 15 sec^{-1}, or higher.

Th us, the k_{cat}/Km of the mutant AARS is optimally equal to or higher for the amino acid analog than for the natural wild type amino acid. Typical k_{cat}/Km values may range from approximately .0001 $M^{-1} s^{-1}$, .0003 $M^{-1} s^{-1}$, .005 $M^{-1} s^{-1}$, .05 $M^{-1} s^{-1}$, .5 $M^{-1} s^{-1}$, .547 $M^{-1} s^{-1}$, 1 $M^{-1} s^{-1}$, 5 $M^{-1} s^{-1}$, 10 $M^{-1} s^{-1}$, 20 $M^{-1} s^{-1}$, 30 $M^{-1} s^{-1}$, 32 $M^{-1} s^{-1}$, 500 $M^{-1}s^{-1}$, 600 $M^{-1} s^{-1}$, 1000 $M^{-1} s^{-1}$, 5000 $M^{-1} s^{-1}$, 11000 $M^{-1} s^{-1}$.

While the point mutations of a mutated AARS typically relate to the binding pocket, the amino acids of the AARS selected for mutation may be altered to any amino acid that allows for aminoacylation of the corresponding tRNA (and thus allows for incorporation of the non-natural amino acid into the target polypeptide or protein).

In certain embodiments, the AARS point mutations may be altered to any amino acid, depending on the characteristics of the non-natural amino acid desired for incorporation into the test protein/polypeptide. In certain embodiments, an amino acid in the binding pocket of the AARS may be mutated to a codon for an amino acid with a small side chain, an amino acid with an aliphatic side chain, a cyclic amino acid, an amino acid with hydroxyl or sulfur containing side chains, an aromatic amino acid, a basic amino acid, an acidic amino acid (or amide). Selection of the amino acid for mutating the AARS at a particular point is routine, depending on the desired outcome and desired non-natural amino acid to be incorporated into the target or test polypeptide/protein. For example, if the goal is to enlarge the binding pocket of

the AARS molecule, then an amino acid with an aliphatic side chain, or a small side chain, could be chosen for mutating the AARS. In other instances, if a binding pocket is desired that harbors a charged pocket, then a basic or acidic amino acid may be selected for point mutation of the AARS.

Certain embodiments disclosed herein include any modified RS molecule in which the binding pocket region has been mutated by at least one point mutation. In certain embodiments, the point mutation are located at one or more positions at which the RS contacts the amino acid for which the RS aminoacylates, or charges, a tRNA molecule. In certain embodiments, multiple point mutations comprise multiple positions at which the RS contacts an amino acid. That is, in certain embodiments multiple or every codon of the entire binding pocket region of an RS may be mutated or modified, or one, two, three, four, or more codons of the binding pocket region of the RS may be mutated or modified. In certain other embodiments, each codon that represents a structural binding point between the particular RS and an amino acid may be mutated or modified. As disclosed herein, multiple different RS molecules have been modified or mutated from various species at the binding point, and guidance is provided for methods that allow one of skill in the art to predictably mutate or modify other RS homolog molecules in the same manner. Certain embodiments provided herein would enable modification and/or mutation of the binding points of the homologous RS molecules in a similar way. Accordingly, such mutation or modification of other RS molecules would be routine experimentation in light of the guidance provided herein.

In some embodiments, the modified RS may be used in a translation system, including an auxotrophic host cell or prototropic host cell along with a suppressor tRNA in order to enable the assignment of a stop codon (such as an amber, ochre or opal nonsense codon, a stop codon that is not present in a particular organism, any nonsense codon, a four or five base pair codon, or another natural amino acid that is not present in significant levels in the protein, such as methionine) to incorporate another amino acid, including

an amino acid analog. Thus, the RS enzymes can be "reprogrammed" for promiscuous substrate specificity in order to facilitate incorporation of a non-natural amino acid into a polypeptide in a site-specific manner. In particular embodiments, any aromatic non-natural amino acid may be utilized with the modified PheRS or TrpRS. This reprogramming allows for high fidelity incorporation of an amino acid, including non-natural amino acids, into polypeptides or proteins with or without the use of auxotrophic host cells.

Reprogramming an AARS enzyme may involve structural or biochemical analysis, including computer modeling or sequence alignment. As there is sequence information available for many AARS molecules, for example at GenBank, comparing sequence alignments is a routine procedure once the particular sequence region of interest is determined.

The use of auxotrophic host cells may increase the level of incorporation of the non-natural amino acid, or decrease the level of misincorporation of another amino acid rather than the desired non-natural amino acid. For example, after enhancing the cellular aminoacylation reactivity by expression of wild type AARS in the host, we surprisingly found that some of the sluggish amino acid analogs could also be introduced into proteins even in the absence of an auxotrophic host cell.

In certain embodiments, the expression of one or more modified or mutant AARS molecules, one or more modified or mutant tRNA molecules, or both, may be regulated by a constitutive or inducible promoter or other inducible expression system.

Certain embodiments disclosed herein relate to allowing for site-selective insertion of one or more unnatural amino acids at any desired position of any protein, (ii) is applicable to both prokaryotic and eukaryotic cells, and enables *in vivo* studies of mutant proteins in addition to the generation of large quantities of purified mutant proteins. In addition, certain embodiments relate to adapting to incorporate any of a large variety of unnatural amino acids, into proteins *in vivo*. Thus, in a specific polypeptide sequence a number of different

site-selective insertions of unnatural amino acids is possible. Such insertions are optionally all of the same type (*e.g.*, multiple examples of one type of unnatural amino acid inserted at multiple points in a polypeptide) or are optionally of diverse types (*e.g.*, different unnatural amino acid types are inserted at multiple points in a polypeptide).

One surprising result disclosed herein shows that the re-design of the synthetic site of an AARS enzyme can expand the ability to introduce replacement amino acids (including non-natural amino acids) into polypeptides or proteins. In some embodiments, the compositions and methods of modifying or mutating an AARS may optionally include altering the editing function of the modified or mutant AARS. In some embodiments, the editing or proofreading ability of the modified or mutant AARS is approximately equal to that of the wild type (unaltered) AARS. In other embodiments, the editing or proofreading ability of the modified or mutant AARS is reduced. In still other embodiments, the editing or proofreading ability of the modified or mutant AARS is eliminated. In some certain embodiments, the alteration of the AARS' editing or proofreading function is inherent to modification of the AARS in order to accommodate a replacement amino acid (for example, modification to the binding pocket of the AARS). In other embodiments, the alteration to the editing or proofreading function may be performed in addition to the modification of the AARS in order to accommodate a replacement amino acid. In still other embodiments, the editing or proofreading function of the AARS is unaltered. Thus, in addition to modification or alteration of the binding pocket of an AARS, the proofreading or editing domain of the modified AARS may also be altered in order to allow for increased specificity of aminoacylating the replacement amino acid (including non-natural amino acid) to a tRNA (whether endogenous or external mutant tRNA), while optionally hydrolyzing the wild type amino acid(s)-adenylate that may form and resulting in greater fidelity or specificity of incorporation of the replacement amino acid (including a non-natural amino acid) into the polypeptide or protein.

In some embodiments, the incorporation rates of a non-natural amino acid were approximately 65% or greater, 70% or greater, 75% or greater, 80% or greater, 85% or greater, 90% or greater, 91% or greater, 92% or greater, 93% or greater, 94% or greater, 95% or greater, 96% or greater, 97% or greater, 98% or greater, or 99% or greater utilizing a modified RS.

As disclosed herein inter alia, the crystal structure of the exact RS to be modified, or the crystal structure of a homologous RS can be used for molecular modeling of the enzyme—amino acid interaction in order to determine the contact points between the RS and the corresponding naturally occurring amino acid, and/or the contact points between the RS and the selected non-natural amino acid desired to be incorporated into a polypeptide. For example, in certain embodiments herein, the crystal structure of *Thermus thermophilus* PheRS complexed with phenylalanine was used for the molecular modeling design of a *Saccharomyces cerevisiae* PheRS, due to the sequence identity of approximately 40% in the active site region of the synthetases. Mutation of the Threonine at position 415 to Glycine or Alanine (T415G or T415A, respectively) enlarged the active site and enabled accommodation of larger phenylalanine analogs. Mutations such as this that disrupt the Watson-Crick base pairing with the naturally occurring amino acid designated for a particular RS allow for increased specificity for incorporation of a non-natural amino acid and decreased misincorporation of another amino acid. As set forth in the Examples and Figures, the (T415A) yeast phenylalanine aminoacyl tRNA synthetase (PheRS) revealed a 5-fold preference for bromophenylalanine than for naturally occurring phenylalanine. Thus, it is possible to alter an aminoacyl tRNA synthetase molecule (RS) to preferentially incorporate a desired amino acid at 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, or greater, depending on the properties of the particular RS and desired amino acid (including non-natural amino acid).

In another particular embodiment, the TrpRS (tryptophan aminoacyl tRNA synthetase) active site generally accommodates bulky non-

natural amino acids, other specific point mutations may allow for further specificity with regard to amino acid incorporation. For example, the point mutation of the D at position 132 of the TrpRS may be altered to a hydrophobic amino acid, which allows for incorporation of bulky non-natural amino acids (such as phenylalanine-derived amino acids). Further mutations at other particular positions in the binding site at locations where recognition occurs for special functional groups of non-natural amino acids may allow for increased specificity and/or higher fidelity of incorporation of the desired amino acid (including non-natural amino acids).

In one particular embodiment, it was found that the space around the para position of the aryl ring of bound Phe could be slightly reduced to exclude other aromatic amino acids, such as Trp, while still accommodating a non-natural amino acid, such as pBrF. Thus, one embodiment discloses incorporation of an aryl bromide functional group into a polypeptide at a programmed position by providing a chemoselective ligation via palladium catalyzed cross-coupling with ethyne or acetylene reaction partners. While such reprogrammed or modified RSs may be used in any number of host cells, including auxotrophic host cells, the high level of efficiency of incorporation of the desired amino acid (or analog) of the modified RSs of certain embodiments, as well as high yields of protein production, render the use of auxotrophic host cells unnecessary.

As the active site region for almost all amino acid synthetases is known or readily deduced, such an exemplary technique may be applied to other AARSs in an effort to reprogram the amino acid specificity from a naturally occurring amino acid to a non-natural amino acid with an expectation of success and without undue experimentation.

As an illustrative example, the threonine at position 415 in yeast PheRS is the equivalent to threonine 251 in *E.coli* PheRS. Thus, mutation of the yeast PheRS (T415G) allowed for activation of a variety of Phe analogs. (See Examples, herein). Further point mutations and/or use of an auxotrophic

host cell allowed for decreased misincorporation in the T415G yeast PheRS variant.

In another particular embodiment disclosed herein, a mutant yeast transfer RNA ($ytRNA^{Phe}_{CUA}$) of which a Watson-Crick base pairing between amino acid position 30 and amino acid position 40, was disrupted was charged with p-bromo-phenylalanine (pBrF) by a co-expressed yeast phenylalanine. In certain embodiments, the amino acid binding pocket of the AARS constitutes approximately 200, approximately 100, approximately 75, approximately 50, approximately 25, approximately 10, approximately 5 or more or less amino acids. In some embodiments, the amino acids to be mutated in the active or binding site are contiguous stretches of amino acids. In other embodiments, the amino acids to be mutated are located within a close proximity to each other but are not contiguous.

In certain embodiments, the natural amino acid is encoded by two or more genetic codes (thus encoded by *degenerate* genetic codes). In most, if not all cases, this includes 18 of the 20 natural amino acids, except Met and Trp. In these circumstances, to recognize all the degenerate genetic codes for the natural amino acid, the anticodon loop of the wild-type tRNA(s) relies on both wobble base-pairing and pure Watson-Crick base-pairing. The subject modified tRNA contains at least one modification in its anticodon loop, such that the modified anticodon loop now forms Watson-Crick base-pairing to one of the degenerate genetic codes, which the tRNA previously bind only through wobble base-pairing.

Since Watson-Crick base pairing is invariably stronger and more stable than wobble base pairing, the subject modified tRNA will preferentially bind to a previous wobble base-pairing genetic code (now through Watson-Crick base-pairing), over a previous Watson-Crick base-pairing (now through wobble base-pairing). Thus an analog may be incorporated at the subject codon, if the modified tRNA is charged with an analog of a natural amino acid,

which may or may not be the same as the natural amino acid encoded by the codon in question.

Thus in certain embodiments, if it is desirable to incorporate certain amino acid analogs at codons for Met or Trp, a tRNA for a natural amino acid (*e.g.*, a Met tRNA, a Trp tRNA, or even a Phe tRNA, etc.) may be modified to recognize the Met or Trp codon. Under this type of unique situation, both the modified tRNA and the natural tRNA compete to bind the same (single) genetic code through Watson-Crick base-pairing. Some, but not all such codons will accept their natural amino acids, while others may accept amino acid analogs carried by the modified tRNA. Other factors, such as the abundance of the natural amino acid vs. that of the analog, may affect the final outcome. (See Examples disclosed herein).

In certain preferred embodiments, the modified tRNA is not charged or only inefficiently charged by an endogenous aminoacyl-tRNA synthetase (AARS) for any natural amino acid, such that the modified tRNA largely (if not exclusively) carries an amino acid analog, but not a natural amino acid. Although a subject modified tRNA may still be useful if it can be charged by the endogenous AARS with a natural amino acid.

In certain embodiments, the modified tRNA charged with an amino acid analog has such an overall shape and size that the analog-tRNA is a ribosomally acceptable complex, that is, the tRNA-analog complex can be accepted by the prokaryotic or eukaryotic ribosomes in an *in vivo* or *in vitro* translation system.

Preferably, the modified AARS specifically or preferentially charges the analog to the modified tRNA over any natural amino acid. In a preferred embodiment, the specificity constant for activation of the analog by the modified AARS (defined as k_{cat} / K_M) is equal to or greater than at least about 2-fold larger than that for the natural amino acid, preferably about 3-fold, 4-fold, 5-fold, 6 fold, 7 fold, 8 fold, 9 fold, 10 fold or more than that for the natural amino acid.

In certain embodiments, the modified tRNA further comprises a mutation at the fourth, extended anticodon site for increase translational efficiency.

The use of extended codons is based on frameshift suppression of translation. Four base codons have the potential for insertion of multiple non-natural amino acids into the same protein. For example, the quadruplet UAGA can be decoded by a $tRNA^{Leu}$ with a UCUA anticodon with an efficiency of 13 to 26%. (See, for example, Moore, *et al.*, *J. Mol. Biol.*, 298: 195 (2000)). The use of extended codons alone has potential problems, such as in-frame readthrough of the first three bases as a triplet in the extended codon competes with the overall frameshift suppression. In some cases, extended codons based on rare codons or nonsense codons may reduce missense readthrough and frameshift suppression at other undesired sites. These problems may be overcome, however, with the use of an extended codon/anticodon and a modified AARS and/or tRNA as indicated in some embodiments disclosed herein.

Thus, to summarize, specific codons are reserved for use in methods disclosed herein by the mutant or modified AARS and/or modified or mutant tRNA for incorporation of a replacement amino acid (including a naturally occurring or non-natural amino acid). Such methods may include use of amber (ochre, umber, or other suppressor tRNA) decoding that reads stop (TAG) codons, bias decoding that exploits unused tRNAs responsible for codon bias, wobble decoding, that creates new tRNAs that read wobble codons, and extended (4-5 base or more) codons that use mutant "suppressor" tRNAs that use 4 base or 5 base (or more) anticodons.

At least one other embodiment provides a method for incorporating an amino acid analog into a target protein at one or more specified positions, the method comprising: (1) providing to an environment a first subject polynucleotide for a modified tRNA, or a subject modified tRNA; (2) providing to the environment a second subject polynucleotide encoding a

modified AARS, wherein the modified AARS is capable of charging the modified tRNA with the analog; (3) providing to the environment the analog; (4) providing a template polynucleotide encoding the target protein, wherein the codon on the template polynucleotide for the specified position only forms Watson-Crick base-pairing with the modified tRNA; and, (5) allowing translation of the template polynucleotide to proceed, thereby incorporating the analog into the target protein at the specified position, wherein steps (1)-(4) are effectuated in any order.

In certain embodiments, the method further comprises verifying the incorporation of the analog by, for example, mass spectrometry, protein sequencing, amino acid tagging such as by fluorescence, radioactivity, etc., ELISA, or other antibody screening, functional assays or screenings, or other methods.

In certain embodiments, the method incorporates the analog into the position at an efficiency of at least about 50%, or 60%, 70%, 80%, 90%, 95%, 99% or nearly 100%.

Definitions

Before describing certain embodiments in detail, it is to be understood that this invention is not limited to particular compositions or biological systems, which can, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular illustrative embodiments only, and is not intended to be limiting. As used in this specification and the appended claims, the singular forms "a," "an," and "the" include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to "a molecule" optionally includes a combination of two or more such molecules, and the like.

Unless specifically defined below, the terms used in this specification generally have their ordinary meanings in the art, within the general context of this invention and in the specific context where each term is

used. Certain terms are discussed below or elsewhere in the specification, to provide additional guidance to the practitioner in describing the compositions and methods of the invention and how to make and use them. The scope and meaning of any use of a term will be apparent from the specific context in which the term is used.

"About" and "approximately" shall generally mean an acceptable degree of error for the quantity measured given the nature or precision of the measurements. Typical, exemplary degrees of error are within 20 percent (%), preferably within 10%, and more preferably within 5% of a given value or range of values. Alternatively, and particularly in biological systems, the terms "about" and "approximately" may mean values that are within an order of magnitude, preferably within 5-fold and more preferably within 2-fold of a given value. Numerical quantities given herein are approximate unless stated otherwise, meaning that the term "about" or "approximately" can be inferred when not expressly stated.

"Amino acid analog," "non-canonical amino acid," or "non-standard amino acid," "non-natural amino acid," "unnatural amino acid," and the like may all be used interchangeably, and is meant to include all amino acid-like compounds that are similar in structure and/or overall shape to one or more of the twenty L-amino acids commonly found in naturally occurring proteins (Ala or A, Cys or C, Asp or D, Glu or E, Phe or F, Gly or G, His or H, Ile or I, Lys or K, Leu or L, Met or M, Asn or N, Pro or P, Gln or Q, Arg or R, Ser or S, Thr or T, Val or V, Trp or W, Tyr or Y, as defined and listed in WIPO Standard ST.25 (1998), Appendix 2, Table 3). Amino acid analog can also be natural amino acids with modified side chains or backbones. Amino acids can also be naturally occurring amino acids in D-, rather than L- form. Preferably, these analogs usually are not "substrates" for the aminoacyl tRNA synthethases (AARSs) because of the normally high specificity of the AARSs. Although occasionally, certain analogs with structures or shapes sufficiently close to those of natural amino acids may be erroneously incorporated into proteins by

AARSs, especially modified AARSs with relaxed substrate specificity. In a preferred embodiment, the analogs share backbone structures, and/or even the most side chain structures of one or more natural amino acids, with the only difference(s) being containing one or more modified groups in the molecule. Such modification may include, without limitation, substitution of an atom (such as N) for a related atom (such as S), addition of a group (such as methyl, or hydroxyl group, etc.) or an atom (such as Cl or Br, etc.), deletion of a group (supra), substitution of a covalent bond (single bond for double bond, etc.), or combinations thereof. Amino acid analogs may include α-hydroxy acids, and β-amino acids, and can also be referred to as "modified amino acids," or "unnatural AARS substrates."

The amino acid analogs may either be naturally occurring or non-natural (*e.g.*, synthesized). As will be appreciated by those in the art, any structure for which a set of rotamers is known or can be generated can be used as an amino acid analog. The side chains may be in either the (R) or the (S) configuration (or D- or L- configuration). In a preferred embodiment, the amino acids are in the (S) or L-configuration.

Preferably, the overall shape and size of the amino acid analogs are such that, upon being charged to (natural or modified or re-designed) tRNAs by (natural or re-designed) AARS, the analog-tRNA is a ribosomally accepted complex, *i.e.*, the tRNA-analog complex can be accepted by the prokaryotic or eukaryotic ribosomes in an *in vivo* or *in vitro* translation system.

"Achor residues" are residue positions in AARS that maintain critical interactions between the AARS and the natural amino acid backbone.

"Backbone," or "template" includes the backbone atoms and any fixed side chains (such as the anchor residue side chains) of the protein (*e.g.*, AARS). For calculation purposes, the backbone of an analog is treated as part of the AARS backbone.

"Protein backbone structure" or grammatical equivalents herein is meant the three dimensional coordinates that define the three dimensional

structure of a particular protein. The structures which comprise a protein backbone structure (of a naturally occurring protein) are the nitrogen, the carbonyl carbon, the α-carbon, and the carbonyl oxygen, along with the direction of the vector from the α-carbon to the β-carbon.

The protein backbone structure that is input into a computer for computational molecular structural or interaction prediction, can either include the coordinates for both the backbone and the amino acid side chains, or just the backbone, *i.e.*, with the coordinates for the amino acid side chains removed. If the former is done, the side chain atoms of each amino acid of the protein structure may be "stripped" or removed from the structure of a protein, as is known in the art, leaving only the coordinates for the "backbone" atoms (the nitrogen, carbonyl carbon and oxygen, and the α-carbon, and the hydrogen atoms attached to the nitrogen and α-carbon).

Optionally, the protein backbone structure may be altered prior to the analysis outlined below. In this embodiment, the representation of the starting protein backbone structure is reduced to a description of the spatial arrangement of its secondary structural elements. The relative positions of the secondary structural elements are defined by a set of parameters called supersecondary structure parameters. These parameters are assigned values that can be systematically or randomly varied to alter the arrangement of the secondary structure elements to introduce explicit backbone flexibility. The atomic coordinates of the backbone are then changed to reflect the altered supersecondary structural parameters, and these new coordinates are input into the system for use in the subsequent protein design automation. For details, see U.S. Pat. No. 6,269,312, the entire content incorporated herein by reference.

"Conformational energy" refers generally to the energy associated with a particular "conformation", or three-dimensional structure, of a macromolecule, such as the energy associated with the conformation of a particular protein. Interactions that tend to stabilize a protein have energies that

are represented as negative energy values, whereas interactions that destabilize a protein have positive energy values. Thus, the conformational energy for any stable protein is quantitatively represented by a negative conformational energy value. Generally, the conformational energy for a particular protein will be related to that protein's stability. In particular, molecules that have a lower (*i.e.*, more negative) conformational energy are typically more stable, *e.g.*, at higher temperatures (*i.e.*, they have greater "thermal stability"). Accordingly, the conformational energy of a protein may also be referred to as the "stabilization energy."

Typically, the conformational energy is calculated using an energy "force-field" that calculates or estimates the energy contribution from various interactions which depend upon the conformation of a molecule. The force-field is comprised of terms that include the conformational energy of the alpha-carbon backbone, side chain - backbone interactions, and side chain - side chain interactions. Typically, interactions with the backbone or side chain include terms for bond rotation, bond torsion, and bond length. The backbone-side chain and side chain-side chain interactions include van der Waals interactions, hydrogen-bonding, electrostatics and solvation terms. Electrostatic interactions may include Coulombic interactions, dipole interactions and quadrapole interactions). Other similar terms may also be included. Force-fields that may be used to determine the conformational energy for a polymer are well known in the art and include the CHARMM (see, Brooks et al, *J. Comp. Chem.* 1983,4:187-217; MacKerell *et al.*, in *The Encyclopedia of Computational Chemistry*, Vol. 1:271-277, John Wiley & Sons, Chichester, 1998), AMBER (see, Cornell *et al.*, *J. Amer. Chem. Soc.* 1995, 117:5179; Woods *et al.*, *J. Phys. Chem.* 1995, 99:3832-3846; Weiner *et al.*, *J. Comp. Chem.* 1986, 7:230; and Weiner *et al.*, *J. Amer. Chem. Soc.* 1984, 106:765) and DREIDING (Mayo *et al.*, *J. Phys. Chem.* 1990, 94-:8897) force-fields, to name but a few.

In a preferred implementation, the hydrogen bonding and electrostatics terms are as described in Dahiyat & Mayo, (*Science* 1997 278:82). The force field can also be described to include atomic conformational terms (bond angles, bond lengths, torsions), as in other references. See *e.g.*, Nielsen, et al. *Prot. Eng.*, 12: 657662(1999); Stikoff, et al., *Biophys. J.*, 67: 2251-2260 (1994); Hendsch, et al., *Prot. Sci.*, 3: 211-226 (1994); Schneider, et al., *J. Am. Chem. Soc.*, 119: 5742-5743 (1997); Sidelar, et al., *Prot. Sci.*, 7: 1898-1914 (1998). Solvation terms could also be included. See *e.g.*, Jackson, et al., *Biochemistry*, 32: 11259-11269 (1993); Eisenberg, et al., *Nature*, 319: 199-203 (1986); Street A G and Mayo S L; *Folding & Design*, 3: 253-258 (1998); Eisenberg and Wesson, *Prot. Sci.*, 1: 227-235 (1992); Gordon & Mayo, supra.

"Coupled residues" are residues in a molecule that interact, through any mechanism. The interaction between the two residues is therefore referred to as a "coupling interaction." Coupled residues generally contribute to polymer fitness through the coupling interaction. Typically, the coupling interaction is a physical or chemical interaction, such as an electrostatic interaction, a van der Waals interaction, a hydrogen bonding interaction, or a combination thereof. As a result of the coupling interaction, changing the identity of either residue will affect the "fitness" of the molecule, particularly if the change disrupts the coupling interaction between the two residues. Coupling interaction may also be described by a distance parameter between residues in a molecule. If the residues are within a certain cutoff distance, they are considered interacting.

"Fitness" is used to denote the level or degree to which a particular property or a particular combination of properties for a molecule, *e.g.*, a protein, are optimized. In certain embodiments of the invention, the fitness of a protein is preferably determined by properties which a user wishes to improve. Thus, for example, the fitness of a protein may refer to the protein's thermal stability, catalytic activity, binding affinity, solubility (*e.g.*, in aqueous or

organic solvent), and the like. Other examples of fitness properties include enantioselectivity, activity towards unnatural substrates, and alternative catalytic mechanisms. Coupling interactions can be modeled as a way of evaluating or predicting fitness (stability). Fitness can be determined or evaluated experimentally or theoretically, *e.g.*, computationally.

Preferably, the fitness is quantitated so that each molecule, *e.g.*, each amino acid will have a particular "fitness value". For example, the fitness of a protein may be the rate at which the protein catalyzes a particular chemical reaction, or the protein's binding affinity for a ligand. In a particularly preferred embodiment, the fitness of a protein refers to the conformational energy of the polymer and is calculated, *e.g.*, using any method known in the art. See, *e.g.*, Brooks, et al., *J. Comp. Chem.*, 4: 187-217 (1983); Mayo, et al., *J. Phys. Chem.*, 94: 8897-8909 (1990); Pabo, et al., *Biochemistry*, 25: 5987-5991 (1986), Lazar, et al., *Prot. Sci.*, 6: 1167-1178 (1997); Lee, et al., *Nature*, 352: 448-451 (1991); Colombo, et al., *J. Am. Chem. Soc.*, 121: 6895-6903 (1999); Weiner, et al., *J. Am. Chem. Soc.*, 106: 765-784 (1984). Generally, the fitness of a protein is quantitated so that the fitness value increases as the property or combination of properties is optimized. For example, in embodiments where the thermal stability of a protein is to be optimized (conformational energy is preferably decreased), the fitness value may be the negative conformationl energy; *i.e.*, F=−E.

The "fitness contribution" of a protein residue refers to the level or extent $f(i_a)$ to which the residue i_a, having an identity a, contributes to the total fitness of the protein. Thus, for example, if changing or mutating a particular amino acid residue will greatly decrease the protein's fitness, that residue is said to have a high fitness contribution to the polymer. By contrast, typically some residues i_a in a protein may have a variety of possible identities a without affecting the protein's fitness. Such residues, therefore have a low contribution to the protein fitness.

"Dead-end elimination" (DEE) is a deterministic search algorithm that seeks to systematically eliminate bad rotamers and combinations of rotamers until a single solution remains. For example, amino acid residues can be modeled as rotamers that interact with a fixed backbone. The theoretical basis for DEE provides that, if the DEE search converges, the solution is the global minimum energy conformation (GMEC) with no uncertainty (Desmet *et al.*, 1992).

Dead end elimination is based on the following concept. Consider two rotamers, i_r and i_t, at residue i, and the set of all other rotamer configurations {S} at all residues excluding i (of which rotamer j_s is a member). If the pairwise energy contributed between i_r and j_s is higher than the pairwise energy between i_t and j_s for all {S}, then rotamer i_r cannot exist in the global minimum energy conformation, and can be eliminated. This notion is expressed mathematically by the inequality.

$$E(i_r) + \sum_{j \neq i}^{N} E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} E(i_t, j_s) \{ S \} \qquad \text{(Equation A)}$$

If this expression is true, the single rotamer i_r can be eliminated (Desmet *et al.*, 1992).

In this form, Equation A is not computationally tractable because, to make an elimination, it is required that the entire sequence (rotamer) space be enumerated. To simplify the problem, bounds implied by Equation A can be utilized:

$$E(i_r) + \sum_{j \neq i}^{N} \min(s) E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max(s) E(i_t, j_s) \{ S \} \qquad \text{(Equation B)}$$

Using an analogous argument, Equation B can be extended to the elimination of pairs of rotamers inconsistent with the GMEC. This is done by determining that a pair of rotamers i_r at residue i and j_s at residue j, always contribute higher energies than rotamers i_u and j_v with all possible rotamer combinations {L}. Similar to Equation B, the strict bound of this statement is given by:

$$\varepsilon(i_r, j_s) + \sum_{k \neq i,j}^{N} \min(t)\varepsilon(i_r, j_s, k_t) > \varepsilon(i_u, j_v) + \sum_{k \neq i,j}^{N} \max(t)\varepsilon(i_u, j_v, k_i) \quad \text{(Equation C)}$$

where ε is the combined energies for rotamer pairs

$$\varepsilon(i_r,j_s) = E(i_r) + E(j_s) + E(i_r,j_s) \quad \text{(Equation D)},$$

and

$$\varepsilon(i_r,j_s,k_t) = E(i_r,k_t) + E(j_s,k_t) \quad \text{(Equation E)}.$$

This leads to the doubles elimination of the pair of rotamers i_r and j_s, but does not eliminate the individual rotamers completely as either could exist independently in the GMEC. The doubles elimination step reduces the number of possible pairs (reduces S) that need to be evaluated in the right-hand side of Equation 6, allowing more rotamers to be individually eliminated.

The singles and doubles criteria presented by Desmet *et al.* fail to discover special conditions that lead to the determination of more dead-ending rotamers. For instance, it is possible that the energy contribution of rotamer i_t is always lower than i_r without the maximum of i_t being below the minimum of i_r. To address this problem, Goldstein 1994 presented a modification of the criteria that determines if the energy profiles of two rotamers cross. If they do not, the higher energy rotamer can be determined to be dead-ending. The doubles calculation uses significantly more computational time than the singles calculation. To accelerate the process, other computational methods have been developed to predict the doubles calculations that will be the most productive (Gordon & Mayo, 1998). These kinds of modifications, collectively referred to as fast doubles, significantly improved the speed and effectiveness of DEE.

Several other modifications also enhance DEE. Rotamers from multiple residues can be combined into so-called super-rotamers to prompt further eliminations (Desmet *et al.*, 1994; Goldstein, 1994). This has the advantage of eliminating multiple rotamers in a single step. In addition, it has been shown that "splitting" the conformational space between rotamers improves the efficiency of DEE (Pierce *et al.*, 2000). Splitting handles the

following special case. Consider rotamer i_r. If a rotamer i_{t1} contributes a lower energy than i_r for a portion of the conformational space, and a rotamer i_{t2} has a lower energy than i_r for the remaining fraction, then i_r can be eliminated. This case would not be detected by the less sensitive Desmet or Goldstein criteria. In the preferred implementations as described herein, all of the described enhancements to DEE were used.

For further discussion of these methods see, Goldstein, *Biophysical Journal* 66, 1335-1340 (1994); Desmet, et al., *Nature* 356,539-542 (1992); Desmet, et al., *The Protein Folding Problem and Tertiary Structure Prediction* (Jr., K. M. & Grand, S. L., eds.), pp. 307-337 (Birkhauser, Boston, 1994); De Maeyer, et al., *Folding & Design* 2, 53-66 (1997), Gordon, and Mayo, *J. Comp. Chem.* 19, 1505-1514 (1998); Pierce, et al., *J. Comp. Chem.* 21, 999-1009 (2000).

Another calculation, dubbed SCREAM (Side-Chain Rotamer Energy Analysis Method), may be used. SCREAM enables examination of the mechanism of discreimination against non-cognate amino acids, by calculating the relative bidning energies of the 20 natural amino acids to a particular AARS. (See, for example, McClendon, *et al., Prot. Eng. Design & Select.* 19: 195-203 (2006)).

As a first step, the rotamer energy spectrum is calculated for a single amino acid in an empty backbone, with no other moveable sidechains. Next, starting with the lowest rotamers from the empty backbone, fill in the sidechains but eliminate clashes. For example, place sidechains at every site, estimating the energies of low lying excitations from the empty backbone spectrum and calculate pairwise interactions, eliminating configurations having clashes. Thus, $E_{tot}(A,B) = E_{self}(A) + E_{self}(B) + E_{Int.}(A,B) = E_{self}(A,B)$ and $E_{tot}(A,B,C) = E_{self}(A) + E_{self}(B) + E_{self}(C) + E_{Int}(A,B) + E_{Int}(A,C) + E_{Int}(B,C) = E_{self}(AB) + E_{self}(C) + E_{Int}(A,C) + E_{Int}(B,C) = E_{self}(A,B) + E_{self}(C) + E_{Int}(AB, C)$ establishes the recursive relation. Next, all low lying sidechain excitations must be analyzed until the energy distribution of rotamer energies in the empty

backbone ceases to increase. Briefly, the ground state energy for all residues is evaluated, followed by a set of rotamers with the lowest linear sum energy, and finally the next lowest linear sum energy and so forth. Furthermore, electrostatic interactions must be addressed as the charges polarize the environment to shield the charges and reduce the desired amino acid interaction. Since molecular dynamics' modeling methods don't usually account for polarization, the bias is in favor of salt bridges. In order to overcome this, the residues are neutralized and parameters evaluated again according to DREIDING parameterization.

For DREIDING parameterization, the lost charged-charged and charged-dipole interactions are compensated by introducing hydrogen bond term. This can be done in conjunction with other programs, including CHARMM, as described herein. Thus, using a crystallographic structure from a particular AARS, or homologous AARS from another organism, we can use a program such as SCREAM, and HierDock, or others, to predict the binding conformation and binding energy of each of the 20 natural amino acids in the binding site in the best-binding mode and the activating mode, by ordering calculations according to which amino acids compete for binding to a particular AARS.

In particular, selective binding is first run, which provides the amino acid and the molecule of ATP to bind to the active site of the AARS. This sometimes leads to a conformational change. Next, selective activation of the AARS to catalyze the formation of a covalent bond between the amino acid and the ATP, forms an aminoacyl adenylate complexed with the AARS and removes inorganic pyrophosphate. Third, if misactivation of a non-cognate amino acid occurs, the AARS may hydrolytically cleave the aminoacyl adenylate complex (as pre-transfer proofreading). Finally, if a non-cognate aminoacyl adenylate has survived, the AARS may hydrolytically cleave the aminoacyl-tRNA complex (post-transfer proofreading).

Thus, such computational programs allow for reliable prediction of the likelihood of the natural amino acids that complete to bind and aminoacylation by wild-type or mutant AARS enzymes. Utilizing multiple programs (such as SCREAM and HierDock together) reduce the misincorporation rate and allow for greater predictability in selecting amino acid locations that specifically bind the amino acid.

Still other computer modeling programs include SCAP (Side Chain Amino Acid Prediction Program) (Xiang and Honig, *J. Mol. Biol.* 311, 421-430 (2001)), and SCWRL (Side Chain Replacement With a Rotamer Library), which is useful for adding sidechains to a protein backbone based on the backbone-dependent rotamer library. The SCWRL library provides lists of chi1-chi2-chi3-chi4 values and their relative probabilities for residues at given phi-psi values, and explores these conformations to minimize sidechain-backbone clashes and sidechain-sidechain clashes. (See, for example, Bower, *et al.*, *J. Mol. Biol.*, 267, 1268-1282 (1997)).

The computational predictability is due, in large part, to utilize known nucleic acid and/or amino acid sequences of AARS enzymes. For example, the catalytic domain is conserved across all members of a particular class of AARS enzyme. (O'Donoghue and Luthey-Schulten, *Microbiol. And Mol. Biol. Rev.*: 550-573 (2003); Diaz-Lazcoz, *et al.*, *Mol. Biol. Evol.* 15(11): 1548-1561 (1998); Wang, *et al.*, *Chem. Commun.* 1-11 (2002)).

"Expression system" means a host cell, or cellular components and compatible vector under suitable conditions, *e.g.*, for the expression of a protein coded for by foreign DNA carried by the vector and introduced to the host cell. Common expression systems include *E. coli* host cells and plasmid vectors, insect host cells such as Sf9, Hi5 or S2 cells and Baculovirus vectors, *Drosophila* cells (Schneider cells) and expression systems, and mammalian host cells and vectors.

"Host cell" means any cell of any organism that is selected, modified, transformed, grown or used or manipulated in any way for the

production of a substance by the cell. A host cell may be auxotrophic, that is unable to synthesize at least one particular organic compound required for its maintenance or growth and must obtain the compound from another source, such as its environment or culture media. In addition, an auxotrophic host cell may have single, double, triple, quadruple or more levels of auxotrophy, such that it is unable to synthesize one, two, three, four or more organic compounds necessary for its growth or maintenance, respectively. For example, a host cell may be one that is manipulated to express a particular gene, a DNA or RNA sequence, a protein or an enzyme. Host cells may be cultured *in vitro* or one or more cells in a non-human animal (*e.g.*, a transgenic animal or a transiently transfected animal).

Certain embodiments disclosed herein expressly utilize only a cell-free expression or translation system and not a host cell. Certain other embodiments expressly utilize only an auxotrophic host cell. Still certain other embodiments expressly utilize only a non-auxotrophic host cell, or a prototrophic host cell.

Sequence similarity may be relevant to certain embodiments as they may include steps of comparing sequences to each other, including wild-type sequence to one or more mutants. Such comparisons typically comprise alignments of polymer sequences, *e.g.*, using sequence alignment programs and/or algorithms that are well known in the art (for example, BLAST, FASTA and MEGALIGN, to name a few). The skilled artisan can readily appreciate that, in such alignments, where a mutation contains a residue insertion or deletion, the sequence alignment will introduce a "gap" (typically represented by a dash, "-", or "Δ") in the polymer sequence not containing the inserted or deleted residue.

"Homologous", in all its grammatical forms and spelling variations, refers to the relationship between two molecules (*e.g.*, proteins, tRNAs, nucleic acids) that possess a "common evolutionary origin", including proteins from superfamilies in the same species of organism, as well as homologous proteins

from different species of organism. Such proteins (and their encoding nucleic acids) have sequence and/or structural homology, as reflected by their sequence similarity, whether in terms of percent identity or by the presence of specific residues or motifs and conserved positions. Homologous molecules frequently also share similar or even identical functions.

In some aspects, homologous may include a sequence that is at least 50% homologous, but that presents a homologous structure in three dimensions, i.e., includes a substantially similar surface charge or presentation of hydrophobic groups. Since hydrogen bonds, van der Waals forces and hydrophobic interactions may function to bind an amino acid to the binding pocket of an AARS, manipulation of a structure of the AARS may also alter one or more of these forces.

Thus, as used herein, proteins and/or protein sequences are "homologous" when they are derived, naturally or artificially, from a common ancestral protein or protein sequence. Similarly, nucleic acids and/or nucleic acid sequences are homologous when they are derived, naturally or artificially, from a common ancestral nucleic acid or nucleic acid sequence. For example, any naturally occurring nucleic acid can be modified by any available mutagenesis method to include one or more selector codon. When expressed, this mutagenized nucleic acid encodes a polypeptide comprising one or more unnatural amino acid. The mutation process can, of course, additionally alter one or more standard codon, thereby changing one or more standard amino acid in the resulting mutant protein as well. Homology is generally inferred from sequence similarity between two or more nucleic acids or proteins (or sequences thereof). The precise percentage of similarity between sequences that is useful in establishing homology varies with the nucleic acid and protein at issue, but as little as 25% sequence similarity is routinely used to establish homology. Higher levels of sequence similarity, *e.g.*, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% or 99% or more can also be used to establish homology. Methods for determining sequence similarity percentages (*e.g.*, BLASTP and

BLASTN using default parameters) are described herein and are generally available.

The term "sequence similarity", in all its grammatical forms, refers to the degree of identity or correspondence between nucleic acid or amino acid sequences that may or may not share a common evolutionary origin (see, Reeck *et al.*, supra). However, in common usage and in the instant application, the term "homologous", when modified with an adverb such as "highly", may refer to sequence similarity and may or may not relate to a common evolutionary origin.

A nucleic acid molecule is "hybridizable" to another nucleic acid molecule, such as a cDNA, genomic DNA, or RNA, when a single stranded form of the nucleic acid molecule can anneal to the other nucleic acid molecule under the appropriate conditions of temperature and solution ionic strength (see Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual,* Second Edition (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.). The conditions of temperature and ionic strength determine the "stringency" of the hybridization. For preliminary screening for homologous nucleic acids, low stringency hybridization conditions, corresponding to a T_m (melting temperature) of 55°C, can be used, *e.g.*, 5×SSC, 0.1% SDS, 0.25% milk, and no formamide; or 30% formamide, 5×SSC, 0.5% SDS). Moderate stringency hybridization conditions correspond to a higher T_m, *e.g.*, 40% formamide, with 5× or 6×SSC. High stringency hybridization conditions correspond to the highest T_m, *e.g.*, 50% formamide, 5× or 6×SSC. SSC is a 0.15M NaCl, 0.015M Na-citrate. Hybridization requires that the two nucleic acids contain complementary stretches of genetic or amino acid sequences, although depending on the stringency of the hybridization, mismatches between bases are possible.

The appropriate stringency for hybridizing nucleic acids depends on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of similarity or homology between

two nucleotide sequences, the greater the value of T_m for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA. For hybrids of greater than 100 nucleotides in length, equations for calculating T_m have been derived (see Sambrook *et al.*, supra, 9.50-9.51). For hybridization with shorter nucleic acids, *i.e.*, oligonucleotides, the position of mismatches becomes more important, and the length of the oligonucleotide determines its specificity (see Sambrook *et al.*, supra, 11.7-11.8). A minimum length for a hybridizable nucleic acid is at least about 10 nucleotides; preferably at least about 15 nucleotides; and more preferably the length is at least about 20 nucleotides.

Unless specified, the term "standard hybridization conditions" refers to a T_m of about 55°C, and utilizes conditions as set forth above. In a preferred embodiment, the T_m is 60°C; in a more preferred embodiment, the T_m is 65°C. In a specific embodiment, "high stringency" refers to hybridization and/or washing conditions at 68°C in 0.2×SSC, at 42°C in 50% formamide, 4×SSC, or under conditions that afford levels of hybridization equivalent to those observed under either of these two conditions.

Suitable hybridization conditions for oligonucleotides (*e.g.*, for oligonucleotide probes or primers) are typically somewhat different than for full-length nucleic acids (*e.g.*, full-length cDNA), because of the oligonucleotides' lower melting temperature. Because the melting temperature of oligonucleotides will depend on the length of the oligonucleotide sequences involved, suitable hybridization temperatures will vary depending upon the oligoncucleotide molecules used. Exemplary temperatures may be 37°C (for 14-base oligonucleotides), 48°C (for 17-base oligoncucleotides), 55°C (for 20-base oligonucleotides) and 60°C (for 23-base oligonucleotides). Exemplary suitable hybridization conditions for oligonucleotides include washing in 6×SSC/0.05% sodium pyrophosphate, or other conditions that afford equivalent levels of hybridization.

"Polypeptide," "peptide" or "protein" are used interchangably to describe a chain of amino acids that are linked together by chemical bonds called "peptide bonds." A protein or polypeptide, including an enzyme, may be a "native" or "wild-type", meaning that it occurs in nature; or it may be a "mutant", "variant" or "modified", meaning that it has been made, altered, derived, or is in some way different or changed from a native protein or from another mutant.

"Rotamer" is defined as a set of possible conformers for each amino acid or analog side chain. See Ponder, *et al.*, Acad. Press Inc. (London) Ltd. pp. 775-791 (1987); Dunbrack, *et al.*, *Struc. Biol.* 1(5):334-340 (1994); Desmet, *et al.*, *Nature* 356:539-542 (1992). A "rotamer library" is a collection of a set of possible / allowable rotametic conformations for a given set of amino acids or analogs. There are two general types of rotamer libraries: "backbone dependent" and "backbone independent." A backbone dependent rotamer library allows different rotamers depending on the position of the residue in the backbone; thus for example, certain leucine rotamers are allowed if the position is within an α helix, and different leucine rotamers are allowed if the position is not in an α-helix. A backbone independent rotamer library utilizes all rotamers of an amino acid at every position. In general, a backbone independent library is preferred in the consideration of core residues, since flexibility in the core is important. However, backbone independent libraries are computationally more expensive, and thus for surface and boundary positions, a backbone dependent library is preferred. However, either type of library can be used at any position.

"Variable residue position" herein is meant an amino acid position of the protein to be designed that is not fixed in the design method as a specific residue or rotamer, generally the wild-type residue or rotamer. It should be noted that even if a position is chosen as a variable position, it is possible that certain methods disclosed herein will optimize the sequence in such a way as to select the wild type residue at the variable position. This generally occurs

more frequently for core residues, and less regularly for surface residues. In addition, it is possible to fix residues as non-wild type amino acids as well.

"Fixed residue position" means that the residue identified in the three dimensional structure as being in a set conformation. In some embodiments, a fixed position is left in its original conformation (which may or may not correlate to a specific rotamer of the rotamer library being used). Alternatively, residues may be fixed as a non-wild type residue depending on design needs; for example, when known site-directed mutagenesis techniques have shown that a particular residue is desirable (for example, to eliminate a proteolytic site or alter the substrate specificity of an AARS), the residue may be fixed as a particular amino acid. Residues which can be fixed include, but are not limited to, structurally or biologically functional residues, for example, the anchor residues.

In certain embodiments, a fixed position may be "floated"; the amino acid or analog at that position is fixed, but different rotamers of that amino acid or analog are tested. In this embodiment, the variable residues may be at least one, or anywhere from 0.1% to 99.9% of the total number of residues. Thus, for example, it may be possible to change only a few (or one) residues, or most of the residues, with all possibilities in between.

As used herein, the term "external mutant" refers to a modified molecule (*e.g.*, an external mutant tRNA and/or an external mutant aminoacyl tRNA synthetase) that exhibits a reduced efficiency (as compared to wild-type or endogenous) for aminoacylation with the corresponding wild type amino acid. "External mutant" refers to the inability or reduced efficiency, *e.g.*, less than 20% efficient, less than 10% efficient, less than 5% efficient, or, *e.g.*, less than 1% efficient, of a tRNA and/or RS to function with the corresponding naturally occurring amino acid in the translation system of interest. For example, an external mutant RS in a translation system of interest aminoacylates any endogenous tRNA of a translation system of interest with the wild type amino

acid at reduced or even zero efficiency, when compared to aminoacylation of an endogenous tRNA by the endogenous RS.

It should be noted, however, that an external mutant RS aminoacylates an endogenous tRNA with a replacement amino acid (whether naturally occurring or non-natural) with an increased efficiency compared with the ability of the endogenous RS to aminoacylate an endogenous tRNA with a replacement amino acid. Likewise, an external mutant tRNA functions at a higher efficiency toward the replacement amino acid (whether non-natural or other naturally occurring amino acid) than toward the corresponding wild type amino acid.

"Wobble degenerate codon" refers to a codon encoding a natural amino acid, which codon, when present in mRNA, is recognized by a natural tRNA anticodon through at least one non-Watson-Crick, or wobble base-pairing (*e.g.*, A-C or G-U base-pairing). Watson-Crick base-pairing refers to either the G-C or A-U (RNA or DNA/RNA hybrid) or A-T (DNA) base-pairing. When used in the context of mRNA codon – tRNA anticodon base-pairing, Watson-Crick base-pairing means all codon-anticodon base-pairings are mediated through either G-C or A-U pairings.

The term "preferentially aminoacylates" refers to an efficiency, *e.g.*, about 20%, about 30%, about 40%, about 50%, about 60%, about 70%, about 75%, about 85%, about 90%, about 95%, about 99% or more efficient. The efficiency may be measured by which a modified or external mutant aminoacyl tRNA synthetase aminoacylates a tRNA with a replacement amino acid, whether an unnatural amino acid or another naturally occurring amino acid when compared to the corresponding natural amino acid assigned to the particular tRNA, AARS, or both. The term "preferentially aminoacylates" further may refer to the efficiency of the modified or external mutant aminoacyl tRNA synthetase to aminoacylate or charge a tRNA with any amino acid other than the corresponding natural amino acid assigned to the particular tRNA, AARS, or both. The term "preferentially aminoacylates" further may refer to the efficiency

of the modified or external mutant aminoacyl tRNA synthetase to aminoacylate a tRNA with a non-natural amino acid compared with the non-modified or naturally occurring AARS.

It should be noted that the efficiency of aminoacylation of the tRNA by the AARS may be correlated to the efficiency of specificity, or fidelity of incorporation of the non-natural amino acid in the target polypeptide or protein. This is due to the function of the protein synthesis machinery in that once a tRNA is aminoacylated with an amino acid (whether the wild type amino acid, or a non-natural amino acid), the charged tRNA is released from the AARS enzyme and the amino acid is incorporated into the target polypeptide. When the proofreading ability of the AARS is altered, the enzyme will allow the replacement amino acid to charge the tRNA and be released for incorporation into the target protein. Thus, the efficiency of aminoacylation by the AARS directly correlates to the fidelity or specificity of incorporation of the non-natural amino acid into the target polypeptide.

The replacement (whether non-natural or naturally occurring) amino acid is then incorporated into a growing polypeptide chain with high fidelity, *e.g.*, at greater than about 20%, 30%, 40%, 50%, 60%, 75%, 80%, 90%, 95%, or greater than about 99% efficiency for a particular codon.

The term "complementary" refers to components of an external mutant pair, the external mutant tRNA and external mutant synthetase that can function together, *e.g.*, the external mutant synthetase aminoacylates the external mutant tRNA.

The term "derived from" refers to a component that is isolated from an organism or isolated and modified, or generated, *e.g.*, chemically synthesized, using information of the component from the organism.

The term "translation system" refers to the components necessary to incorporate a naturally occurring or unnatural amino acid into a growing polypeptide chain (protein). For example, components can include ribosomes, tRNA(s), synthetas(es), mRNA and the like. The components disclosed herein

can be added to a translation system, *in vivo* or *in vitro*. An *in vivo* translation system may be a cell (eukaryotic or prokaryotic cell). An *in vitro* translation system may be a cell-free system, such as reconstituted one with components from different organisms (purified or recombinantly produced). In certain embodiments, the translation system does not comprise a cell. In certain embodiments, the translation system does not comprise an auxotrophic cell. If the translation system does not comprise an auxtrophic cell, it may comprise another cell or cellular components.

The term "inactive RS" refers to a synthetase that has been mutated so that it no longer can aminoacylate its cognate tRNA with any amino acid, whether naturally occurring or non-natural. The term "modified RS" refers to a synthetase that has been mutated so that it no longer can aminoacylate its cognate tRNA with the corresponding naturally occurring amino acid, but may be able to aminoacylate its cognate tRNA with another amino acid, preferably a non-natural amino acid.

The term "selection agent" refers to an agent that when present allows for a selection of certain components from a population, *e.g.*, an antibiotic, wavelength of light, an antibody, a nutrient or the like. The selection agent can be varied, *e.g.*, such as concentration, intensity, etc.

The term "positive selection marker" refers to a marker than when present, *e.g.*, expressed, activated or the like, results in identification of an organism with the positive selection marker from those without the positive selection marker.

The term "negative selection marker" refers to a marker than when present, *e.g.*, expressed, activated or the like, allows identification of an organism that does not possess the desired property (*e.g.*, as compared to an organism which does possess the desired property).

The term "reporter" refers to a component that can be used to select components described in the disclosure. For example, a reporter can

include a green fluorescent protein, a firefly luciferase protein, or genes such as β-gal/lacZ (β-galactosidase), Adh (alcohol dehydrogenase) or the like.

The term "not efficiently recognized" refers to an efficiency, *e.g.*, less than about 10%, less than about 5%, or less than about 1%, at which a RS from one organism aminoacylates an external mutant tRNA. In certain embodiments, the RS may be from the same or a different organism than the external mutant tRNA. In some embodiments, the RS has been modified to aminoacylate a tRNA with a particular amino acid, preferably a non-natural amino acid.

The term "eukaryote" refers to organisms belonging to the phylogenetic domain Eucarya such as animals (*e.g.*, mammals, insects, reptiles, birds, etc.), ciliates, plants, fungi (*e.g.*, yeasts, etc.), flagellates, microsporidia, protists, etc. Additionally, the term "prokaryote" refers to non-eukaryotic organisms belonging to the Eubacteria (*e.g.*, *Escherichia coli*, *Thermus thermophilus*, etc.) and Archaea (*e.g.*, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Halobacterium* such as *Haloferax volcanii* and *Halobacterium* species NRC-1, *A. fulgidus, P. firiosus, P. horikoshii, A. pernix*, etc.) phylogenetic domains.

The Genetic Code, Host Cells, and the Degenerate Codons

The standard genetic code most cells use is listed below.

The Genetic Code

Middle

First	U	C	A	G	Last
	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	Stop		
(Ochre)	Stop				
(Umber)	A				

First	U	C	A	G	Last
	Leu	Ser	Stop		
(Amber)	Trp	G			
	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The genetic code is degenerate, in that the protein biosynthetic machinery utilizes 61 mRNA sense codons to direct the templated polymerization of the 20 natural amino acid monomers. (Crick *et al.*, Nature 192: 1227, 1961). Just two amino acids, *i.e.*, methionine and tryptophan, are encoded by unique mRNA triplets.

The standard genetic code applies to most, but not all, cases. Exceptions have been found in the mitochondrial DNA of many organisms and in the nuclear DNA of a few lower organisms. Some examples are given in the following table.

Examples of non-standard genetic codes.

Mitochondria	Vertebrates	UGA→ Trp; AGA, AGG → STOP
	Invertebrates	UGA→ Trp; AGA, AGG → Ser
	Yeasts	UGA→ Trp; CUN → Thr
	Protista	UGA→ Trp;
Nucleus	Bacteria	GUG, UUG, AUU, CUG → initiation
	Yeasts	CUG → Ser
	Ciliates	UAA, UAG → Gln

*Plant cells use the standard genetic code in both mitochondria and the nucleus.

The NCBI (National Center for Biotechnology Information) maintains a detailed list of the standard genetic code, and genetic codes used in various organisms, including the vertebrate mitochondrial code; the yeast mitochondrial code; the mold, protozoan, and coelenterate mitochondrial code and the mycoplasma / spiroplasma code; the invertebrate mitochondrial code; the ciliate, dasycladacean and hexamita nuclear code; the echinoderm and flatworm mitochondrial code; the euplotid nuclear code; the bacterial and plant plastid code; the alternative yeast nuclear code; the ascidian mitochondrial code; the alternative flatworm mitochondrial code; blepharisma nuclear code; chlorophycean mitochondrial code; trematode mitochondrial code; scenedesmus obliquus mitochondrial code; thraustochytrium mitochondrial code (all incorporated herein by reference). These are primarily based on the reviews by Osawa *et al.*, *Microbiol. Rev.* 56: 229-264, 1992, and Jukes and Osawa, *Comp. Biochem. Physiol.* 106B: 489-494, 1993.

Host Cells

Some methods disclosed herein can be practiced within a cell, which enables production levels of proteins to be made for practical purposes. In preferred embodiments, the cells used are culturable cells (*i.e.*, cells that can

be grown under laboratory conditions). Suitable cells include mammalian cells (human or non-human mammals), bacterial cells, and insect cells, etc.

One example includes PFENEX™ technology, which is a cell line using *Pseudomonas fluorescens*-based cell line that increase cellular expression while maintaining certain solubility and activity characteristics due to its use of different pathways in the metabolism of certain sugars compared to *E.coli*.

In addition, other auxotrophic host cell lines include K10 based Phe ausotrophic strain (AF), Phe/Trp double auxotrophic strains (AFW), Phe/Trp/Lys triple auxotrophic strains (AFWK), a Met auxotroph (M15MA on M15 background), as well as DH10B based AF strain.

Cells that may be used to practice certain embodiments disclosed herein include auxotrophic host cells (whether prokaryotic or eukaryotic). Auxotrophic cells may exhibit single, double, triple, quadruple, or greater levels of auxotrophy (each level of auxotrophy indicating a particular organic compound of which the organism is unable to synthesize and must be supplied to the cell). Certain embodiments disclosed herein expressly do not utilize an auxotrophic host cell. Insofar as an auxotrophic host cell is not used, another cell or cell components may still be used to practice certain embodiments disclosed herein. Other embodiments may use one, two, three, or more different auxotrophic host cells that may be from the same or different strains or organisms.

Host cells are genetically engineered (*e.g.*, transformed, transduced or transfected) with the vectors of this disclosure, which can be, for example, a cloning vector or an expression vector. The vector can be, for example, in the form of a plasmid, a bacterium, a virus, a naked polynucleotide, or a conjugated polynucleotide. The vectors are introduced into cells and/or microorganisms by standard methods including electroporation (From *et al.*, *PNAS*. USA 82, 5824 (1985)), infection by viral vectors, high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of

small beads or particles, or on the surface (Klein *et al.*, *Nature* 327, 70-73 (1987)). Berger, Sambrook, and Ausubel provide a variety of appropriate transformation methods.

The engineered host cells can be cultured in conventional nutrient media modified as appropriate for such activities as, for example, screening steps, activating promoters or selecting transfonnants. These cells can optionally be cultured into transgenic organisms.

Certain embodiments disclosed herein further include methods of screening modified AARSs and/or modified tRNAs. For example, in one embodiment, a yeast PheRS library is subjected to double sieve screening in order to detect high levels of incorporation of a non-natural amino acid or misincorporation of natural amino acids other than Phe will lead to severe misfolding or unfolding of GFP. The yeast PheRS library cells are thus subjected to high-throughput screening based on flow cytommetry analysis (FACS). First, the yeast PheRS library cells are expressed in the presence of 2Nal and low fluorescent cells indicating higher incorporation of either 2Nal or other natural amino acids are collected by FACS. Next, the yeast PheRS library cells are expressed without 2Nal. Bright cells are collected in order to eliminate yeast PheRS variants that can misincorporate other natural amino acids. In on exemplary embodiment, two cycles of screening yielded a mutant yeast PheRS with mutations in N412G, S418C, T415G and S437F, which had low fluorescence in the presence of 2Nal and high fluorescence in the absence of 2Nal. This technique allows for incorporation of 2Nal at UUU codon, increasing to around 90%.

Other useful references, *e.g.*, for cell isolation and culture (*e.g.*, for subsequent nucleic acid isolation) include Freshney (1994) *Culture of Animal Cells, a Manual of Basic Technique*, third edition, Wiley-Liss, New York and the references cited therein; Payne *et al.* (1992) *Plant Cell and Tissue Culture in Liquid Systems* John Wiley & Sons, Inc. New York, N.Y.; Gamborg and Phillips (eds.) (1995) *Plant Cell, Tissue and Organ Culture; Fundamental*

Methods Springer Lab Manual, Springer-Verlag (Berlin Heidelberg New York) and Atlas and Parks (eds.) *The Handbook of Microbiological Media* (1993) CRC Press, Boca Raton, Fla.

Several well-known methods of introducing target nucleic acids into bacterial cells are available, any of which can be used in certain embodiments disclosed herein. These include: fusion of the recipient cells with bacterial protoplasts containing the DNA, electroporation, projectile bombardment, and infection with viral vectors, etc. Bacterial cells can be used to amplify the number of plasmids containing DNA constructs of certain embodiments disclosed herein. The bacteria are grown to log phase and the plasmids within the bacteria can be isolated by a variety of methods known in the art (see, for instance, Sambrook). In addition, a plethora of kits are commercially available for the purification of plasmids from bacteria, (see, *e.g.*, EASYPREP™, FLEXIPREP™, both from Pharmacia Biotech; STRATACLEAN™, from Stratagene; and, QIAPREP™ from Qiagen). The isolated and purified plasmids are then further manipulated to produce other plasmids, used to transfect cells or incorporated into related vectors to infect organisms.

Typical vectors contain transcription and translation terminators, transcription and translation initiation sequences, and promoters useful for regulation of the expression of the particular target nucleic acid. The vectors optionally comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the cassette in eukaryotes, or prokaryotes, or both, (*e.g.*, shuttle vectors) and selection markers for both prokaryotic and eukaryotic systems. Vectors are suitable for replication and integration in prokaryotes, eukaryotes, or preferably both. See Giliman & Smith, *Gene* 8:81 (1979); Roberts, *et al.*, *Nature*, 328:731 (1987); Schneider, B., *et al.*, *Protein Expr. Purif.* 6435:10 (1995); Ausubel, Sambrook, Berger (all supra). A catalogue of Bacteria and Bacteriophages useful for cloning is provided, *e.g.*, by the ATCC, *e.g.*, *The ATCC Catalogue of*

Bacteria and Bacteriophage (1992) Gherna *et al.* (eds.) published by the ATCC. Additional basic procedures for sequencing, cloning and other aspects of molecular biology and underlying theoretical considerations are also found in Watson *et al.* (1992) *Recombinant DNA* 2nd Edition Scientific American Books, NY.

Degenerate Codon Selection

As described above, all amino acids, with the exception of methionine and tryptophan are encoded by more than one codon. According to some methods disclosed herein, a codon in the genome that is normally used to encode a natural amino acid is reprogrammed, in part by the transcriptional or translational machinery to instead encode an amino acid analog. An amino acid analog can be a naturally occurring or canonical amino acid analog. In a preferred embodiment, the amino acid analog is not a canonically encoded amino acid.

The thermodynamic stability of a codon-anticodon pair can be predicted or determined experimentally. According to some embodiments, it is preferable that the external mutant tRNA interacts with the degenerate codon with an affinity (at 37°C) of at least about 1.0 kcal/mol more strongly, even more preferably 1.5 kcal/mole more strongly, and even more preferably more than 2.0 kcal/mol more strongly than a natural tRNA in the cell would recognize the same sequence. These values are known to one of skill in the art and can be determined by thermal denaturation experiments (see, *e.g.*, Meroueh and Chow, *Nucleic Acids Res.* 27: 1118, 1999).

The following table lists some of the known anti-codon sequences for *E. coli*. In general, for any organism, tRNA anticodon sequence can be routinely determined using art-recognized technologies. For example, any tRNA gene can be amplified by, for example, PCR. Sequencing can be performed to determine the exact sequences of the anti-codon loop. Alternatively, biochemical binding assay may be used to determine the binding

affinity of a purified tRNA to one of the 2-6 possible codons. The codon that binds the tRNA with the highest specificity / affinity presumably has pure Watson-Crick match at all three codon positions, thus determining the sequence of the anti-codon loop.

In general, the wobble base in the anti-codon loop tends to be G or U (rather than A or C).

The Degenerate Codons for *E. coli*

Amino Acid	Anti-codon	Base-paring at 3rd base	Codon	Amino Acid	Anti-codon	Base-paring	Codon
Ala	GGC	W/C[1]	GCC	His	GUG	W/C	CAC
		Wobble[2]	GCU			Wobble	CAU
	UGC	W/C	GCA	Ile	GAU	W/C	AUC
		Wobble	GCG			Wobble	AUU, AUA
Asp	GUC	W/C	GAC	Leu	GAG	W/C	CUC, CUA, CUG, UUC, UUG
		Wobble	GAU			Wobble	CUU
Asn	GUU	W/C	AAC	Lys	UUU	W/C	AAA
		Wobble	AAU			Wobble	AAG
Cys	GCA	W/C	UGC	Phe	GAA	W/C	UUC
		Wobble	UGU			Wobble	UUU
Glu	UUC	W/C	GAA	Ser	GGA	W/C	UUC, AGU
		Wobble	GAG			Wobble	UCU, AGC, UCA, UCG

Amino Acid	Anti-codon	Base-paring at 3rd base	Codon	Amino Acid	Anti-codon	Base-paring	Codon
Gly	GCC	W/C	GGC, GGA, GGG	Tyr	GUA	W/C	UAC
		Wobble	GGU			Wobble	UAU
Met		W/C	AUG	Thr		W/C	ACC, ACA, ACG
Gln		W/C	CAA, CAG			Wobble	ACU
Arg		W/C	AGA, AGG, CGU, CGG	Pro		W/C	CCC, CCA, CCG
		Wobble	CGC, CGA	Trp		Wobble	CCU
						W/C	UGG
STOP		W/C	UGA, UAA	Val		W/C	GUC, GUA
		Wobble	UAG			Wobble	GUU, GUG

[1] Watson-Crick base pairing
[2] Wobble base pairing

When a single tRNA recognizes a codon through a perfect complementary interaction between the anticodon of the tRNA and one codon, it is called Watson-Crick base pairing. When a single tRNA recognizes a second, degenerate codon, it is called a wobble or other non-standard base pairing. In certain embodiments disclosed herein, a new tRNA can be constructed having an anticodon sequence that is perfectly complementary to a degenerate codon or a codon for a non-natural amino acid, thus utilizing wobble or Watson-Crick base pairing. Likewise, a new AARS can be constructed that utilizes a replacement amino acid (other than wild type—may be another naturally occurring amino acid or a non-natural amino acid) to aminoacylate the

corresponding tRNA. This may be in addition to or instead of modifying a tRNA molecule for incorporation of a replacement amino acid.

The modified AARS may be altered such that the binding efficiency to the non-natural amino acid, or another selected naturally occurring amino acid, is greater than the binding efficiency of the modified AARS to the corresponding naturally occurring amino acid. In this way, a modified AARS may preferentially bind a non-natural amino acid in order to charge a tRNA even in the presence of the naturally occurring amino acid that corresponds to the AARS in its unmodified state. This "reprogramming" of an aminoacyl tRNA synthetase allows for incorporation of a non-natural amino acid into a polypeptide with lower levels of mis-incorporation of other amino acids into the desired site.

The "reprogramming" further may allow for use of the modified or external mutant synthetase with high levels of incorporation in standard host cells, without the need for auxotrophic host cells, and with or without depleting the media of the corresponding naturally occurring amino acid. Thus, while certain embodiments disclosed herein may be practiced by using an auxotrophic host cell, certain other embodiments may be practiced without using an auxotrophic host cell. In the event of not using an auxotrophic host cell to practice certain embodiments, another host cell may be used, cellular components may be used, or an entirely cell-free system may be used.

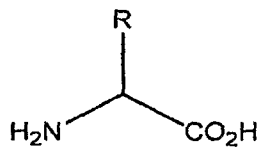
When the cell has multiple tRNA molecules for a particular amino acid, and one tRNA has an anticodon sequence that is perfectly complementary to the degenerate codon selected, the gene encoding the tRNA can be disabled through any means available to one of skill in the art including, for example, site-directed mutagenesis or deletion of either the gene or the promoter sequence of the gene. Expression of the gene also can be disable through any antisense or RNA interference techniques.

Unnatural or Non-natural Amino Acids

The first step in the protein engineering process is usually to select a set of unnatural or non-natural amino acids that have the desired chemical properties. The selection of non-natural amino acids depends on pre-determined chemical properties one would like to have, and the modifications one would like to make in the target protein. Unnatural amino acids, once selected, can either be purchased from vendors, or chemically synthesized.

A wide variety of unnatural or non-natural amino acids can be used in the methods disclosed herein. The unnatural amino acid can be chosen based on desired characteristics of the unnatural amino acid, *e.g.*, function of the unnatural amino acid, such as modifying protein biological properties such as toxicity, biodistribution, immunogenicity, or half life, structural properties, spectroscopic properties, chemical and/or photochemical properties, catalytic properties, ability to react with other molecules (either covalently or noncovalently), or the like.

As used herein an "unnatural amino acid" refers to any amino acid, modified amino acid, or amino acid analogue other than selenocysteine and the following twenty genetically encoded alpha-amino acids: alanine, arginine, asparagine, aspartic acid, cysteine, glutamine, glutamic acid, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine. The generic structure of an alpha-amino acid is illustrated by Formula I:



Formula I

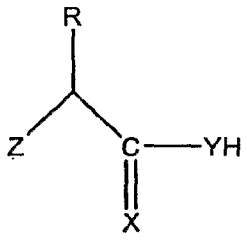
An unnatural amino acid is typically any structure having Formula I wherein the R group is any substituent other than one used in the twenty natural amino acids. See, *e.g.*, any biochemistry text such as Biochemistry by L. Stryer, 3rd ed. 1988, Freeman and Company, New York, for structures of the

twenty natural amino acids. Note that the unnatural amino acids disclosed herein may be naturally occurring compounds other than the twenty alpha-amino acids above. Because the unnatural amino acids disclosed herein typically differ from the natural amino acids in side chain only, the unnatural amino acids form amide bonds with other amino acids, *e.g.*, natural or unnatural, in the same manner in which they are formed in naturally occurring proteins. However, the unnatural amino acids have side chain groups that distinguish them from the natural amino acids. For example, R in Formula I optionally comprises an alkyl-, aryl-, aryl halide, vinyl halide, alkyl halide, acetyl, ketone, aziridine, nitrile, nitro, halide, acyl-, keto-, azido-, hydroxyl-, hydrazine, cyano-, halo-, hydrazide, alkenyl, alkynyl, ether, thioether, epoxide, sulfone, boronic acid, boronate ester, borane, phenylboronic acid, thiol, seleno-, sulfonyl-, borate, boronate, phospho, phosphono, phosphine, heterocyclic-, pyridyl, naphthyl, benzophenone, a constrained ring such as a cyclooctyne, thioester, enone, imine, aldehyde, ester, thioacid, hydroxylamine, amino, carboxylic acid, alpha-keto carboxylic acid, alpha or beta unsaturated acids and amides, glyoxyl amide, or organosilane group, or the like or any combination thereof.

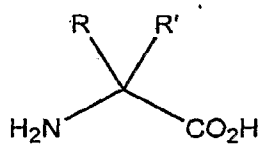
Aryl substitutions may occur at various positions, *e.g.* ortho, meta, para, and with one or more functional groups placed on the aryl ring. Other unnatural amino acids of interest include, but are not limited to, amino acids comprising a photoactivatable cross-linker, spin-labeled amino acids, dye-labeled amino acids, fluorescent amino acids, metal binding amino acids, metal-containing amino acids, radioactive amino acids, amino acids with novel functional groups, amino acids with altered hydrophilicity, hydrophobocity, polarity, or ability to hydrogen bond, amino acids that covalently or noncovalently interact with other molecules, photocaged and/or photoisomerizable amino acids, amino acids comprising biotin or a biotin analogue, glycosylated amino acids such as a sugar substituted serine, other carbohydrate modified amino acids, keto containing amino acids, amino acids

comprising polyethylene glycol or a polyether, a polyalcohol, or a polysaccharide, amino acids that can undergo metathesis, amino acids that can undergo cycloadditions, heavy atom substituted amino acids, chemically cleavable and/or photocleavable amino acids, amino acids with an elongated side chains as compared to natural amino acids, *e.g.*, polyethers or long chain hydrocarbons, *e.g.*, greater than about 5 or greater than about 10 carbons, carbon-linked sugar-containing amino acids, redox-active amino acids, amino thioacid containing amino acids, amino acids containing a drug moiety, and amino acids comprising one or more toxic moieties.

In addition to unnatural amino acids that contain novel side chains, unnatural amino acids also optionally comprise modified backbone structures, *e.g.*, as illustrated by the structures of Formula II and III:



Formula II



Formula III

wherein Z typically comprises OH, NH_2, SH, NH_2O-, NH-R', R'NH-, R'S-, or S-R'-; X and Y, which may be the same or different, typically comprise S, N, or O, and R and R', which are optionally the same or different, are typically selected from the same list of constituents for the R group described above for the unnatural amino acids having Formula I as well as hydrogen or $(CH_2)_x$ or the natural amino acid side chains. For example, unnatural amino acids disclosed herein optionally comprise substitutions in the amino or carboxyl group as illustrated by Formulas II and III. Unnatural amino acids of this type include, but are not limited to, α-hydroxy acids, α-thioacids α-aminothiocarboxylates, or α-α-disubstituted amino acids, with side chains corresponding *e.g.*to the twenty natural amino acids or to unnatural side chains. They also include but are not

limited to β-amino acids or γ-amino acids, such as substituted β-alanine and γ-amino butyric acid. In addition, substitutions or modifications at the α-carbon optionally include L or D isomers, such as D-glutamate, D-alanine, D-methyl-O-tyrosine, aminobutyric acid, and the like. Other structural alternatives include cyclic amino acids, such as proline analogs as well as 3-, 4-, 6-, 7-, 8-, and 9-membered ring proline analogs. Some non-natural amino acids, such as aryl halides (p-bromo-phenylalanine, p-iodophenylalanine, provide versatile palladium catalyzed cross-coupling reactions with ethyne or acetylene reactions that allow for formation of carbon-carbon, carbon-nitrogen and carbon-oxygen bonds between aryl halides and a wide variety of coupling partners.

For example, many unnatural amino acids are based on natural amino acids, such as tyrosine, glutamine, phenylalanine, and the like. Tyrosine analogs include para-substituted tyrosines, ortho-substituted tyrosines, and meta substituted tyrosines, wherein the substituted tyrosine comprises an acetyl group, a benzoyl group, an amino group, a hydrazine, an hydroxyamine, a thiol group, a carboxy group, an isopropyl group, a methyl group, a C6-C20 straight chain or branched hydrocarbon, a saturated or unsaturated hydrocarbon, an O-methyl group, a polyether group, a nitro group, or the like. In addition, multiply substituted aryl rings are also contemplated. Glutamine analogs include, but are not limited to, α-hydroxy derivatives, β-substituted derivatives, cyclic derivatives, and amide substituted glutamine derivatives. Exemplary phenylalanine analogs include, but are not limited to, meta-substituted phenylalanines, wherein the substituent comprises a hydroxy group, a methoxy group, a methyl group, an allyl group, an acetyl group, or the like.

Specific examples of unnatural amino acids include, but are not limited to, *o, m* and/or *p* forms of amino acids or amino acid analogs (non-natural amino acids), including homoallylglycine, cis- or trans-crotylglycine, 6,6,6-trifluoro-2-aminohexanoic acid, 2-aminopheptanoic acid, norvaline, norleucine, O-methyl-L-tyrosine, o-, m-, or p-methyl-phenylalanine, O-4-allyl-L-tyrosine, a 4-propyl-L-tyrosine, a tri-O-acetyl-GlcNAcβ-serine, an L-Dopa, a

fluorinated phenylalanine, an isopropyl-L-phenylalanine, a p-azidophenylalanine, a p-acyl-L-phenylalanine, a p-benzoyl-L-phenylalanine, an L-phosphoserine, a phosphonoserine, a phosphonotyrosine, a p-iodo-phenylalanine, o-, m-, or p-bromophenylalanine, 2-, 3-, or 4-pyridylalanine, p-idiophenylalanine, diaminobutyric acid, aminobutyric acid, benzofuranylalanine, 3-bromo-tyrosine, 3-(6-chloroindolyl)alanine, 3-(6-bromoindolyl)alanine, 3-(5-bromonindolyl)alanine, p-chlorophenylalanine, p-ethynyl-phenylalanine, p-propargly-oxy-phenylalanine, m-ethynyl-phenylalanine, 6-ethynyl-tryptophan, 5-ethynyl-tryptophan, (R)-2-amino-3-(4-ethynyl-1H-pyrol-3-yl)propanoic acid, azidonorleucine, azidohomoalanine, p-acetylphenylalanine, p-amino-L-phenylalanine, homopropargiyglycine, p-ethyl-phenylalanine, p-ethynyl-phenylalanine, p-propargly-oxy-phenylalanine, isopropyl-L-phenylalanine, an 3-(2-naphthyl)alanine, 3-(1-naphthyl)alanine, 3-idio-tyrosine, O-propargyl-tyrosine, homoglutamine, an O-4-allyl-L-tyrosine, a 4-propyl-L-tyrosine, a 3-nitro-L-tyrosine, a tri-O-acetyl-GlcNAcβ-serine, an L-Dopa, a fluorinated phenylalanine, an isopropyl-L-phenylalanine, a p-azido-L-phenylalanine, a p-acyl-L-phenylalanine, a p-acetyl-L-phenylalanine, an m-acetyl-L-phenylalanine, selenomethionine, telluromethionine, selenocysteine, an alkyne phenylalanine, an O-allyl-L-tyrosine, an O-(2-propynyl)-L-tyrosine, a p-ethylthiocarbonyl-L-phenylalanine, a p-(3-oxobutanoyl)-L-phenylalanine, a p-benzoyl-L-phenylalanine, an L-phosphoserine, a phosphonoserine, a phosphonotyrosine, homopropargiyglycine, azidohomoalanine, a p-iodo-phenylalanine, a p-bromo-L-phenylalanine, dihydroxy-phenylalanine, dihydroxyl-L-phenylalanine, a p-nitro-L-phenylalanine, an m-methoxy-L-phenylalanine, a p-iodo-phenylalanine, a p-bromophenylalanine, a p-amino-L-phenylalanine, and an isopropyl-L-phenylalanine, trifluoroleucine, norleucine, 4-, 5-, or 6- fluoro-tryptophan, 4-aminotryptophan, 5-hydroxytryptophan, biocytin, aminooxyacetic acid, m-hydroxyphenylalanine, m-allyl phenylalanine, m-methoxyphenylalanine group, β-GlcNAc-serine, α-GalNAc-threonine, p-acetoacetylphenylalanine, para-halo-phenylalanine, seleno-methionine, ethionine, S-nitroso-homocysteine, thia-

proline, 3-thienyl-alanine, homo-allyl-glycine, trifluoroisoleucine, trans and cis-2-amino-4-hexenoic acid, 2-butynyl-glycine, allyl-glycine, para-azido-phenylalanine, para-cyano-phenylalanine, para-ethynyl-phenylalanine, hexafluoroleucine, 1,2,4-triazole-3-alanine, 2-fluoro-histidine, L-methyl histidine, 3-methyl-L-histidine, β-2-thienyl-L-alanine, β-(2-thiazolyl)-DL-alanine, homopropargylglycine (HPG) and azidohomoalanine (AHA) and the like. The structures of a variety of non-limiting unnatural amino acids are provided in the figures, *e.g.*, FIGS. 29, 30, and 31 of US 2003/0108885 A1, the entire content of which is incorporated herein by reference.

Tyrosine analogs include para-substituted tyrosines, ortho-substituted tyrosines, and meta substituted tyrosines, wherein the substituted tyrosine comprises an acetyl group, a benzoyl group, an amino group, a hydrazine, an hydroxyamine, a thiol group, a carboxy group, an isopropyl group, a methyl group, a C6-C20 straight chain or branched hydrocarbon, a saturated or unsaturated hydrocarbon, an O-methyl group, a polyether group, a nitro group, or the like. In addition, multiply substituted aryl rings are also contemplated. Glutamine analogs of the invention include, but are not limited to, α-hydroxy derivatives, β-substituted derivatives, cyclic derivatives, and amide substituted glutamine derivatives. Example phenylalanine analogs include, but are not limited to, meta-substituted phenylalanines, wherein the substituent comprises a hydroxy group, a methoxy group, a methyl group, an allyl group, an acetyl group, or the like.

Additionally, other examples optionally include (but are not limited to) an unnatural analog of a tyrosine amino acid; an unnatural analog of a glutamine amino acid; an unnatural analog of a phenylalanine amino acid; an unnatural analog of a serine amino acid; an unnatural analog of a threonine amino acid; an alkyl, aryl, acyl, azido, cyano, halo, hydrazine, hydrazide, hydroxyl, alkenyl, alkynl, ether, thiol, sulfonyl, seleno, ester, thioacid, borate, boronate, phospho, phosphono, phosphine, heterocyclic, enone, imine, aldehyde, hydroxylamine, keto, or amino substituted amino acid, or any

combination thereof; an amino acid with a photoactivatable cross-linker; a spin-labeled amino acid; a fluorescent amino acid; an amino acid with a novel functional group; an amino acid that covalently or noncovalently interacts with another molecule; a metal binding amino acid; a metal-containing amino acid; a radioactive amino acid; a photocaged amino acid; a photoisomerizable amino acid; a biotin or biotin-analog containing amino acid; a glycosylated or carbohydrate modified amino acid; a keto containing amino acid; an amino acid comprising polyethylene glycol; an amino acid comprising polyether; a heavy atom substituted amino acid; a chemically cleavable or photocleavable amino acid; an amino acid with an elongated side chain; an amino acid containing a toxic group; a sugar substituted amino acid, *e.g.*, a sugar substituted serine or the like; a carbon-linked sugar-containing amino acid; a redox-active amino acid; an α-hydroxy containing acid; an amino thio acid containing amino acid; an α,α disubstituted amino acid; a β-amino acid; and a cyclic amino acid.

Typically, the unnatural amino acids utilized herein for certain embodiments may be selected or designed to provide additional characteristics unavailable in the twenty natural amino acids. For example, unnatural amino acid are optionally designed or selected to modify the biological properties of a protein, *e.g.*, into which they are incorporated. For example, the following properties are optionally modified by inclusion of an unnatural amino acid into a protein: toxicity, biodistribution, solubility, stability, *e.g.*, thermal, hydrolytic, oxidative, resistance to enzymatic degradation, and the like, facility of purification and processing, structural properties, spectroscopic properties, chemical and/or photochemical properties, catalytic activity, redox potential, half-life, ability to react with other molecules, *e.g.*, covalently or noncovalently, and the like.

Other examples of amino acid analogs optionally include (but are not limited to) an unnatural analog of a tyrosine amino acid; an unnatural analog of a glutamine amino acid; an unnatural analog of a phenylalanine amino acid; an unnatural analog of a serine amino acid; an unnatural analog of

a threonine amino acid; an alkyl, aryl, acyl, azido, cyano, halo, hydrazine, hydrazide, hydroxyl, alkenyl, alkynl, ether, thiol, sulfonyl, seleno, ester, thioacid, borate, boronate, phospho, phosphono, phosphine, heterocyclic, enone, imine, aldehyde, hydroxylamine, keto, or amino substituted amino acid, or any combination thereof; an amino acid with a photoactivatable cross-linker; a spin-labeled amino acid; a fluorescent amino acid; an amino acid with a novel functional group; an amino acid that covalently or noncovalently interacts with another molecule; a metal binding amino acid; a metal-containing amino acid; a radioactive amino acid; a photocaged amino acid; a photoisomerizable amino acid; a biotin or biotin-analogue containing amino acid; a glycosylated or carbohydrate modified amino acid; a keto containing amino acid; an amino acid comprising polyethylene glycol; an amino acid comprising polyether; a heavy atom substituted amino acid; a chemically cleavable or photocleavable amino acid; an amino acid with an elongated side chain; an amino acid containing a toxic group; a sugar substituted amino acid, *e.g.*, a sugar substituted serine or the like; a carbon-linked sugar-containing amino acid; a redox-active amino acid; an α-hydroxy containing acid; an amino thio acid containing amino acid; an α,α disubstituted amino acid; a β-amino acid; and a cyclic amino acid other than proline.

Aminoacyl-tRNA Synthetases

The aminoacyl-tRNA synthetase (used interchangeably herein with AARS, RS or "synthetase") used in certain methods disclosed herein can be a naturally occurring synthetase derived from an organism, whether the same (homologous) or different (heterologous), a mutated or modified synthetase, or a designed synthetase.

The synthetase used can recognize the desired (unnatural) amino acid analog selectively over related amino acids available. For example, when the amino acid analog to be used is structurally related to a naturally occurring amino acid, the synthetase should charge the external mutant tRNA molecule

with the desired amino acid analog with an efficiency at least substantially equivalent to that of, and more preferably at least about twice, 3 times, 4 times, 5 times or more than that of the naturally occurring amino acid. However, in cases in which a well-defined protein product is not necessary, the synthetase can have relaxed specificity for charging amino acids. In such an embodiment, a mixture of external mutant tRNAs could be produced, with various amino acids or analogs.

In certain embodiments, it is preferable that the synthetase has activity both for the amino acid analog and for the amino acid that is encoded by the corresponding codon of the tRNA molecule.

A synthetase can be obtained by a variety of techniques known to one of skill in the art, including combinations of such techniques as, for example, computational methods, selection methods, and incorporation of synthetases from other organisms (see below).

In certain embodiments, synthetases can be used or developed that efficiently charge tRNA molecules that are not charged by synthetases of the host cell. For example, suitable pairs may be generally developed through modification of synthetases from organisms distinct from the host cell. In certain embodiments, the synthetase can be developed by selection procedures. In certain embodiments, the synthetase can be designed using computational techniques such as those described in Datta *et al.*, *J. Am. Chem. Soc.* 124: 5652-5653, 2002, and in U.S. Patent No. 7,139,665, hereby incorporated by reference .

Computational Design of AARS

Specifically, in one embodiment, the subject method partly depends on the design and engineering of natural AARS to a modified form that has relaxed substrate specificity, such that it can uptake non-canonical amino acid analogs as a substrate, and charge a modified tRNA (with its anticodon changed) with such a non-canonical amino acid. The following sections briefly

describe a method for the generation of such modified AARS, which method is described in more detail in U.S. Patent No. 7,139,665, the entire contents of which are incorporated herein by reference.

Briefly, the methods of some embodiments described therein relate to computational tools for modifying the substrate specificity of an AminoAcyl tRNA Synthetases (AARSs) through mutation to enable the enzyme to more efficiently utilize amino acid analog(s) in protein translation systems, either *in vitro* , in whole cells, or in other translation systems. A feature of some embodiments includes systematically redesigning the substrate binding site of an AARS enzyme to facilitate the use of unnatural substrates in the peptide or protein translation reaction the enzyme catalyzes.

According to one method, a rotamer library for the artificial amino acid is built by varying its torsional angles to create rotamers that would fit in the binding pocket for the natural substrate. The geometric orientation of the backbone of the amino acid analog is specified by the crystallographic orientation of the backbone of the natural substrate in the crystal structure. The crystallographic structure of the organism-specific amino acid synthetase may be used, or a homologous structure from another organism may be used, depending on structural similarity. Amino acids in the binding pocket of the synthetase that interact with the side chain on the analog are allowed to vary in identity and rotameric conformation in the subsequent protein design calculations.

One such protocol also employs a computational method to enhance the interactions between the substrate and the protein positions. This is done by scaling up the pair-wise energies between the substrate and the amino acids allowed at the design positions on the protein in the energy calculations. In an optimization calculation where the protein-substrate interactions are scaled up compared to the intra-protein interactions, sequence selection is biased toward selecting amino acids to be those that have favorable interaction with the substrate.

The described method helped to construct a new modified form of the *E. coli* phenylalanyl-tRNA synthetase, based on the known structure of the related *Thermus thermophilus* PheRS (*t*PheRS). The new modified form of the *E. coli* phenylalanyl-tRNA synthetase (ePheRS) allows efficient *in vivo* incorporation of reactive aryl ketone functionality into recombinant proteins. In addition, a modified tryptophanyl-tRNA synthetase was modified in a similar manner and has demonstrated the ability to incorporate non-natural amino acid analogs in polypeptides in place of naturally occurring tryptophan. The results described therein also demonstrate the general power of computational protein design in the development of aminoacyl-tRNA synthetases for activation and charging of non-natural amino acids.

A. Available Sequence and Structural Information for tRNA Synthetases

Protein translation from an mRNA template is carried out by ribosomes. During the translation process, each tRNA is matched with its amino acid long before it reaches the ribosome. The match is made by a collection of enzymes known as the aminoacyl-tRNA synthetases (AARS). These enzymes charge each tRNA with the proper amino acid, thus allowing each tRNA to make the proper translation from the genetic code of DNA (and the mRNA transcribed from the DNA) into the amino acid code of proteins.

Most cells make twenty different aminoacyl-tRNA synthetases, one for each type of amino acid. These twenty enzymes are each optimized for function with its own particular amino acid and the set of tRNA molecules appropriate to that amino acid. Aminoacyl-tRNA synthetases must perform their tasks with high accuracy. Many of these enzymes recognize their tRNA molecules using the anticodon. These enzymes make about one mistake in 10,000. For most amino acids, this level of accuracy is not too difficult to achieve, since most of the amino acids are quite different from one another.

In the subject method, an accurate description of the AARS binding pocket for tRNA is important for the computational design approach, since it depends on the crystal structure for the protein backbone descriptions, although in many cases it is perfectly acceptable to use crystal structure of a homologous protein (for example, a homolog from a related species) or even a conserved domain to substitute the crystallographic binding pocket structure description. The crystal structure also defines the orientation of the natural substrate amino acid in the binding pocket of a synthetase, as well as the relative position of the amino acid substrate to the synthetase residues, especially those residues in and around the binding pocket. To design the binding pocket for the analogs, it is preferred that these analogs bind to the synthetase in the same orientation as the natural substrate amino acid, since this orientation may be important for the adenylation step.

The AARSs may be from any organism, including prokaryotes and eukaryotes, with enzymes from bacteria, fungi, extremeophiles such as the archebacteria, worm, insects, fish, amphibian, birds, animals (particularly mammals and particularly human) and plants all possible.

As described above, most cells make twenty different aminoacyl-tRNA synthetases, one for each type of amino acid. Some suitable synthetases are known, including: yeast phenylalanyl-tRNA synthetase (Kwon *et al.*, *J. Am. Chem. Soc.* 125: 7512-7513, 2003); *Methonococcus jannaschii* tyrosyl-tRNA synthetase (Wang *et al.*, *Science* 292, 498-500, 2001); and yeast tyrosyl-tRNA synthetase (Ohno *et al.*, *J. Biochem.* 130, 417-423, 2001). In fact, the crystal structures of nearly all 20 different AARS enzymes are currently available in the Brookhaven Protein Data Bank (PDB, see Bernstein *et al.*, *J. Mol. Biol.* 112: 535-542, 1977). A list of all the AARSs with solved crystal structures as of April 2001 is available on the PDB website. For example, the crystal structure of *Thermus Aquaticus* Phenylalanyl tRNA Synthetase complexed with Phenylalanine has a resolution of 2.7 Å, and its PDB ID is 1B70.

The structure database or Molecular Modeling DataBase (MMDB) contains experimental data from crystallographic and NMR structure determinations. The data for MMDB are obtained from the Protein Data Bank (PDB). The NCBI (National Center for Biotechnology Information) has cross-linked structural data to bibliographic information, to the sequence databases, and to the NCBI taxonomy. Cn3D, the NCBI 3D structure viewer, can be used for easy interactive visualization of molecular structures from Entrez.

The Entrz 3D Domains database contains protein domains from the NCBI Conserved Domain Database (CDD). Computational biologists define conserved domains based on recurring sequence patterns or motifs. CDD currently contains domains derived from two popular collections, Smart and Pfam, plus contributions from colleagues at NCBI, such as COG. The source databases also provide descriptions and links to citations. Since conserved domains correspond to compact structural units, CDs contain links to 3D-structure via Cn3D whenever possible.

To identify conserved domains in a protein sequence, the CD-Search service employs the reverse position-specific BLAST algorithm. The query sequence is compared to a position-specific score matrix prepared from the underlying conserved domain alignment. Hits may be displayed as a pairwise alignment of the query sequence with a representative domain sequence, or as a multiple alignment. CD-Search now is run by default in parallel with protein BLAST searches. While the user waits for the BLAST queue to further process the request, the domain architecture of the query may already be studied. In addition, CDART, the Conserved Domain Architecture Retrieval Tool allows user to search for proteins with similar domain architectures. CDART uses precomputed CD-search results to quickly identify proteins with a set of domains similar to that of the query. For more details, see Marchler-Bauer *et al.*, *Nucleic Acids Research* 31: 383-387, 2003; and Marchler-Bauer *et al.*, *Nucleic Acids Research* 30: 281-283, 2002.

In addition, a database of known aminoacyl tRNA synthetases has been published by Maciej Szymanski, Marzanna A. Deniziak and Jan Barciszewski, in *Nucleic Acids Res.* 29:288-290, 2001 (titled "Aminoacyl-tRNA synthetases database"). A corresponding website (rose.man.poznan.pl/aars/seq_main.html) provides details about all known AARSs from different species. For example, according to the database, the Isoleucyl-tRNA Synthetase for the radioresistant bacteria *Deinococcus radiodurans* (Accession No. AAF10907) has 1078 amino acids, and was published by White *et al.* in *Science* 286:1571-1577(1999); the Valyl-tRNA Synthetase for mouse (*Mus musculus*) has 1263 amino acids (Accession No. AAD26531), and was published by Snoek M. and van Vugt H. in *Immunogenetics* 49: 468-470(1999); and the Phenylalanyl-tRNA Synthetase sequences for human, *Drosophila*, *S. pombe*, *S. cerevisiae*, *Candida albicans*, *E. coli*, and mumerous other bacteria including *Thermus aquaticus ssp. thermophilus* are also available. The database was last updated in November, 2006. Similar information for other newly identified AARSs can be obtained, for example, by conducting a BLAST search using any of the known sequences in the AARS database as query against the available public (such as the non-redundant database at NCBI, or "nr") or proprietory private databases.

Alternatively, in certain embodiments, if the exact crystal structure of a particular AARS is not known, but its protein sequence is similar or homologous to a known AARS sequence with a known crystal structure. In such instances, it is expected that the conformation of the AARS in question will be similar to the known crystal structure of the homologous AARS. The known structure may, therefore, be used as the structure for the AARS of interest, or more preferably, may be used to predict the structure of the AARS of interest (*i.e.*, in "homology modeling" or "molecular modeling"). As a particular example, the Molecular Modeling Database (MMDB) described above (see, Wang *et al.*, *Nucl. Acids Res.* 2000, 28:243-245; Marchler-Bauer *et al.*, *Nucl. Acids Res.* 1999,27:240-243) provides search engines that may be used to identify

proteins and/or nucleic acids that are similar or homologous to a protein sequence (referred to as "neighboring" sequences in the MMDB), including neighboring sequences whose three-dimensional structures are known. The database further provides links to the known structures along with alignment and visualization tools, such as Cn3D (developed by NCBI), RasMol, etc., whereby the homologous and parent sequences may be compared and a structure may be obtained for the parent sequence based on such sequence alignments and known structures.

The homologous AARS sequence with known 3D-structure is preferably at least about 60%, or at least about 70%, or at least about 80%, or at least about 90%, or at least about 95% identical, or at least about 98% identical to the AARS of interest in the active site region or the pocket region for amino acid substrate binding. Such active site or pocket site may not be continuous in the primary amino acid sequence of the AARS since distant amino acids may come together in the 3D-structure. In this case, sequence homology or identity can be calculated using, for example, the NCBI standard BLASTp programs for protein using default conditions, in regions aligned together (without insertions or deletions in either of the two sequences being compared) and including residues known to be involved in substrate amino acid binding. For example, the *Thermus Aquaticus* phenylalanyl tRNA synthetase catalytic (alpha) subunit appears to have an "insert" region from residues 156 to 165 when compared to its homologs from other species. This region can be disregarded in calculating sequence identity. Alternatively, the homologous AARS is preferably about 35%, or 40%, or 45%, or 50%, or 55% identical overall to the AARS of interest. The *E. coli* phenylalanyl tRNA synthetase alpha subunit is about 45% identical overall, and about 80% identical in the active site region to the *Thermus Aquaticus* phenylalanyl tRNA synthetase. The human phenylalanyl tRNA synthetase alpha subunits is about 62%, 60%, 54%, 50% identical overall to its *Drosophila*, worm (*C. elegans*), plant (*Arabidopsis thaliana*), yeast (*S. cerevisiae*) counterparts, respectively.

In the few cases where the structure for a particular AARS sequence may not be known or available, it is possible to determine the structure using routine experimental techniques (for example, X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy) and without undue experimentation. See, *e.g.*, *NMR of Macromolecules: A Practical Approach*, G. C. K. Roberts, Ed., Oxford University Press Inc., New York (1993); Ishima and Torchia, *Nat. Struct. Biol.* 7: 740-743, 2000; Gardner and Kay, *Annu. Rev. Bioph. Biom.* 27: 357-406, 1998; Kay, *Biochem. Cell. Biol.* 75: 1-15, 1997; Dayie *et al.*, *Annu. Rev. Phys. Chem.* 47: 243-282, 1996; Wuthrich, *Acta Cyrstallogr. D* 51: 249-270, 1995; Kahn *et al.*, *J. Synchrotron Radiat.* 7: 131-138, 2000; Oakley and Wilce, *Clin. Exp. Pharmacol. P.* 27: 145-151, 2000; Fourme *et al.*, *J. Synchrotron Radiat.* 6: 834-844, 1999.

Alternatively, the three-dimensional structure of a AARS sequence may be calculated from the sequence itself and using *ab initio* molecular modeling techniques already known in the art. See *e.g.*, Smith *et al.*, *J. Comput. Biol.* 4: 217-225, 1997; Eisenhaber *et al.*, *Proteins* 24: 169-179, 1996; Bohm, *Biophys Chem.* 59: 1-32, 1996; Fetrow and Bryant, *BioTechnol.* 11: 479-484, 1993; Swindells and Thorton, *Curr. Opin. Biotech.* 2: 512-519, 1991; Levitt *et al.*, *Annu. Rev. Biochem.* 66: 549-579, 1997; Eisenhaber *et al.*, *Crit. Rev. Biochem. Mol.* 30: 1-94, 1995; Xia *et al.*, *J. Mol. Biol.* 300: 171-185, 2000; Jones, *Curr. Opin. Struc. Biol.* 10: 371-379, 2000. Three-dimensional structures obtained from *ab initio* modeling are typically less reliable than structures obtained using empirical (*e.g.*, NMR spectroscopy or X-ray crystallography) or semi-empirical (*e.g.*, homology modeling) techniques. However, such structures will generally be of sufficient quality, although less preferred, for use in some methods disclosed herein.

B. Methods for Predicting 3D Structure based on Sequence Homology

For AARS proteins that have not been crystallized or been the focus of other structural determinations, a computer-generated molecular model of the AARS and its binding site can nevertheless be generated using any of a number of techniques available in the art. For example, the Cα-carbon positions of the target AARS sequence can be mapped to a particular coordinate pattern of an AARS enzyme ("known AARS") having a similar sequence and deduced structure using homology modeling techniques, and the structure of the target protein and velocities of each atom calculated at a simulation temperature (To) at which a docking simulation with an amino acid analog is to be determined. Typically, such a protocol involves primarily the prediction of side-chain conformations in the modeled target AARS protein, while assuming a main-chain trace taken from a tertiary structure, such as provided by the known AARS protein. Computer programs for performing energy minimization routines are commonly used to generate molecular models. For example, both the CHARMM (Brooks *et al.* (1983) *J Comput Chem* 4:187-217) and AMBER (Weiner et al (1981) *J. Comput. Chem.* 106: 765) algorithms handle all of the molecular system setup, force field calculation, and analysis (see also, Eisenfield *et al.* (1991) *Am J Physiol* 261:C376-386; Lybrand (1991) *J Pharm Belg* 46:49-54; Froimowitz (1990) *Biotechniques* 8:640-644; Burbam *et al.* (1990) *Proteins* 7:99-111; Pedersen (1985) *Environ Health Perspect* 61:185-190; and Kini *et al.* (1991) *J Biomol Struct Dyn* 9:475-488). At the heart of these programs is a set of subroutines that, given the position of every atom in the model, calculate the total potential energy of the system and the force on each atom. These programs may utilize a starting set of atomic coordinates, the parameters for the various terms of the potential energy function, and a description of the molecular topology (the covalent structure). Common features of such molecular modeling methods include: provisions for handling hydrogen bonds and other constraint forces; the use of

periodic boundary conditions; and provisions for occasionally adjusting positions, velocities, or other parameters in order to maintain or change temperature, pressure, volume, forces of constraint, or other externally controlled conditions.

Most conventional energy minimization methods use the input coordinate data and the fact that the potential energy function is an explicit, differentiable function of Cartesian coordinates, to calculate the potential energy and its gradient (which gives the force on each atom) for any set of atomic positions. This information can be used to generate a new set of coordinates in an effort to reduce the total potential energy and, by repeating this process over and over, to optimize the molecular structure under a given set of external conditions. These energy minimization methods are routinely applied to molecules similar to the subject AARS proteins.

In general, energy minimization methods can be carried out for a given temperature, Ti, which may be different than the docking simulation temperature, To. Upon energy minimization of the molecule at Ti, coordinates and velocities of all the atoms in the system are computed. Additionally, the normal modes of the system are calculated. It will be appreciated by those skilled in the art that each normal mode is a collective, periodic motion, with all parts of the system moving in phase with each other, and that the motion of the molecule is the superposition of all normal modes. For a given temperature, the mean square amplitude of motion in a particular mode is inversely proportional to the effective force constant for that mode. In this regard, the low frequency vibrations will often dominate the motion of the molecule.

After the molecular model has been energy minimized at Ti, the system is "heated" or "cooled" to the simulation temperature, To, by carrying out an equilibration run where the velocities of the atoms are scaled in a step-wise manner until the desired temperature, To, is reached. The system is further equilibrated for a specified period of time until certain properties of the system,

such as average kinetic energy, remain constant. The coordinates and velocities of each atom are then obtained from the equilibrated system.

Further energy minimization routines can also be carried out. For example, a second class of methods involves calculating approximate solutions to the constrained EOM for the protein. These methods use an iterative approach to solve for the Lagrange multipliers and, typically, only need a few iterations if the corrections required are small. The most popular method of this type, SHAKE (Ryckaert *et al.* (1977) *J. Comput. Phys.* 23:327; and Van Gunsteren *et al.* (1977) *Mol. Phys.* 34:1311) is easy to implement and scales as O(N) as the number of constraints increases. Therefore, the method is applicable to macromolecules such as AARS proteins. An alternative method, RATTLE (Anderson (1983) *J. Comput. Phys.* 52:24) is based on the velocity version of the Verlet algorithm. Like SHAKE, RATTLE is an iterative algorithm and can be used to energy minimize the model of a subject AARS protein.

C. Alternative Methods

In other embodiments, rather than holding the identity of the amino acid analog constant and varying the AARS structure (by modeling several different mutant structures), the subject method is carried out using the molecular model(s) for a single modified AARS (*e.g.*, in which one more non-anchor amino acid residues are changed) and sampling a variety of different amino acid analogs or potential fragments thereof, to identify analogs which are likely to interact with, and be substrates for the modified AARS enzyme. This approach can make use of coordinate libraries for amino acid analogs (including rotamer variants) or libraries of functional groups and spacers that can be joined to form the side-chain of an amino acid analog.

Using such approaches as described above, *e.g.*, homology modeling, a coordinate set for the binding site for the modified AARS can be derived.

There are a variety of computational methods that can be readily adapted for identifying the structure of amino acid analogs that would have appropriate steric and electronic properties to interact with the substrate binding site of a modified AARS. See, for example, Cohen *et al.* (1990) *J. Med. Cam.* 33: 883-894; Kuntz *et al.* (1982) *J. Mol. Biol* 161: 269-288; DesJarlais (1988) *J. Med. Cam.* 31: 722-729; Bartlett *et al.* (1989) (*Spec. Publ., Roy. Soc. Chem.*) 78: 182-196; Goodford *et al.* (1985) *J. Med. Cam.* 28: 849-857; DesJarlais *et al. J. Med. Cam.* 29: 2149-2153). Directed methods generally fall into two categories: (1) design by analogy in which 3-D structures of known molecules (such as from a crystallographic database) are docked to the AARS binding site structure and scored for goodness-of-fit; and (2) *de novo* design, in which the amino acid analog model is constructed piece-wise in the AARS binding site. The latter approach, in particular, can facilitate the development of novel molecules, uniquely designed to bind to the subject modified AARS binding site.

In an illustrative embodiment, the design of potential amino acid analogs that may function with a particular modified AARS begins from the general perspective of shape complimentary for the substrate binding site of the enzyme, and a search algorithm is employed which is capable of scanning a database of small molecules of known three-dimensional structure for candidates which fit geometrically into the substrate binding site. Such libraries can be general small molecule libraries, or can be libraries directed to amino acid analogs or small molecules which can be used to create amino acid analogs. It is not expected that the molecules found in the shape search will necessarily be leads themselves, since no evaluation of chemical interaction necessarily be made during the initial search. Rather, it is anticipated that such candidates might act as the framework for further design, providing molecular skeletons to which appropriate atomic replacements can be made. Of course, the chemical complimentary of these molecules can be evaluated, but it is expected that atom types will be changed to maximize the electrostatic,

hydrogen bonding, and hydrophobic interactions with the substrate binding site. Most algorithms of this type provide a method for finding a wide assortment of chemical structures that may be complementary to the shape of the AARS substrate binding site.

For instance, each of a set of small molecules from a particular data-base, such as the Cambridge Crystallographic Data Bank (CCDB) (Allen *et al.* (1973) *J. Chem. Doc.* 13: 119), is individually docked to the binding site of the modified AARS in a number of geometrically permissible orientations with use of a docking algorithm. In a preferred embodiment, a set of computer algorithms called DOCK, can be used to characterize the shape of invaginations and grooves that form the binding site. See, for example, Kuntz *et al.* (1982) *J. Mol. Biol* 161: 269-288. The program can also search a database of small molecules for templates whose shapes are complementary to particular binding site of the modified AARS. Exemplary algorithms that can be adapted for this purpose are described in, for example, DesJarlais *et al.* (1988) *J. Med. Chem.* 31:722-729.

The orientations are evaluated for goodness-of-fit and the best are kept for further examination using molecular mechanics programs, such as AMBER or CHARMM. Such algorithms have previously proven successful in finding a variety of molecules that are complementary in shape to a given binding site of a receptor or enzyme, and have been shown to have several attractive features. First, such algorithms can retrieve a remarkable diversity of molecular architectures. Second, the best structures have, in previous applications to other proteins, demonstrated impressive shape complementarity over an extended surface area. Third, the overall approach appears to be quite robust with respect to small uncertainties in positioning of the candidate atoms.

In certain embodiments, the subject method can utilize an algorithm described by Goodford (1985, *J. Med. Chem.* 28:849-857) and Boobbyer *et al.* (1989, *J. Med. Chem.* 32:1083-1094). Those papers describe a computer program (GRID) which seeks to determine regions of high affinity for

different chemical groups (termed probes) on the molecular surface of the binding site. GRID hence provides a tool for suggesting modifications to known ligands that might enhance binding. It may be anticipated that some of the sites discerned by GRID as regions of high affinity correspond to "pharmacophoric patterns" determined inferentially from a series of known ligands. As used herein, a pharmacophoric pattern is a geometric arrangement of features of the anticipated amino acid analog that is believed to be important for binding. Goodsell and Olson (1990, *Proteins: Struct. Funct. Genet.* 8:195-202) have used the Metropolis (simulated annealing) algorithm to dock a single known ligand into a target protein, and their approach can be adapted for identifying suitable amino acid analogs for docking with the AARS binding site. This algorithm can allow torsional flexibility in the amino acid side-chain and use GRID interaction energy maps as rapid lookup tables for computing approximate interaction energies.

Yet a further embodiment utilizes a computer algorithm such as CLIX which searches such databases as CCDB for small molecules which can be oriented in the substrate binding site of the AARS in a way that is both sterically acceptable and has a high likelihood of achieving favorable chemical interactions between the candidate molecule and the surrounding amino acid residues. The method is based on characterizing the substrate binding site in terms of an ensemble of favorable binding positions for different chemical groups and then searching for orientations of the candidate molecules that cause maximum spatial coincidence of individual candidate chemical groups with members of the ensemble. The current availability of computer power dictates that a computer-based search for novel ligands follows a breadth-first strategy. A breadth-first strategy aims to reduce progressively the size of the potential candidate search space by the application of increasingly stringent criteria, as opposed to a depth-first strategy wherein a maximally detailed analysis of one candidate is performed before proceeding to the next. CLIX conforms to this strategy in that its analysis of binding is rudimentary -it seeks

to satisfy the necessary conditions of steric fit and of having individual groups in "correct" places for bonding, without imposing the sufficient condition that favorable bonding interactions actually occur. A ranked "shortlist" of molecules, in their favored orientations, is produced which can then be examined on a molecule-by-molecule basis, using computer graphics and more sophisticated molecular modeling techniques. CLIX is also capable of suggesting changes to the substituent chemical groups of the candidate molecules that might enhance binding. Again, the starting library can be of amino acid analogs or of molecules which can be used to generate the side-chain of an amino acid analog.

The algorithmic details of CLIX is described in Lawerence *et al.* (1992) *Proteins* 12:31-41, and the CLIX algorithm can be summarized as follows. The GRID program is used to determine discrete favorable interaction positions (termed target sites) in the binding site of the AARS protein for a wide variety of representative chemical groups. For each candidate ligand in the CCDB an exhaustive attempt is made to make coincident, in a spatial sense in the binding site of the protein, a pair of the candidate's substituent chemical groups with a pair of corresponding favorable interaction sites proposed by GRID. All possible combinations of pairs of ligand groups with pairs of GRID sites are considered during this procedure. Upon locating such coincidence, the program rotates the candidate ligand about the two pairs of groups and checks for steric hindrance and coincidence of other candidate atomic groups with appropriate target sites. Particular candidate/orientation combinations that are good geometric fits in the binding site and show sufficient coincidence of atomic groups with GRID sites are retained.

Consistent with the breadth-first strategy, this approach involves simplifying assumptions. Rigid protein and small molecule geometry is maintained throughout. As a first approximation rigid geometry is acceptable as the energy minimized coordinates of the binding site of the modified AARS, describe an energy minimum for the molecule, albeit a local one.

A further assumption implicit in CLIX is that the potential ligand, when introduced into the substrate binding site of the modified AARS, does not induce change in the protein's stereochemistry or partial charge distribution and so alter the basis on which the GRID interaction energy maps were computed. It must also be stressed that the interaction sites predicted by GRID are used in a positional and type sense only, *i.e.*, when a candidate atomic group is placed at a site predicted as favorable by GRID, no check is made to ensure that the bond geometry, the state of protonation, or the partial charge distribution favors a strong interaction between the protein and that group. Such detailed analysis should form part of more advanced modeling of candidates identified in the CLIX shortlist.

Yet another embodiment of a computer-assisted molecular design method for identifying amino acid analogs that may be utilized by a predetermined modified AARS comprises the *de novo* synthesis of potential inhibitors by algorithmic connection of small molecular fragments that will exhibit the desired structural and electrostatic complementarity with the substrate binding site of the enzyme. The methodology employs a large template set of small molecules with are iteratively pieced together in a model of the AARS' substrate binding site. Each stage of ligand growth is evaluated according to a molecular mechanics-based energy function, which considers van der Waals and coulombic interactions, internal strain energy of the lengthening ligand, and desolvation of both ligand and enzyme. The search space can be managed by use of a data tree which is kept under control by pruning according to the binding criteria.

In yet another embodiment, potential amino acid analogs can be determined using a method based on an energy minimization-quenched molecular dynamics algorithm for determining energetically favorable positions of functional groups in the substrate binding site of a modified AARS enzyme. The method can aid in the design of molecules that incorporate such functional

groups by modification of known amino acid and amino acid analogs or through *de novo* synthesis.

For example, the multiple copy simultaneous search method (MCSS) described by Miranker *et al.* (1991) *Proteins* 11: 29-34 can be adapted for use in the subject method. To determine and characterize a local minima of a functional group in the force field of the protein, multiple copies of selected functional groups are first distributed in a binding site of interest on the AARS protein. Energy minimization of these copies by molecular mechanics or quenched dynamics yields the distinct local minima. The neighborhood of these minima can then be explored by a grid search or by constrained minimization. In one embodiment, the MCSS method uses the classical time dependent Hartee (TDH) approximation to simultaneously minimize or quench many identical groups in the force field of the protein.

Implementation of the MCSS algorithm requires a choice of functional groups and a molecular mechanics model for each of them. Groups must be simple enough to be easily characterized and manipulated (3-6 atoms, few or no dihedral degrees of freedom), yet complex enough to approximate the steric and electrostatic interactions that the functional group would have in substrate binding to the site of the AARS protein. A preferred set is, for example, one in which most organic molecules can be described as a collection of such groups (*Patai's Guide to the Chemistry of Functional Groups*, ed. S. Patai (New York: John Wiley, and Sons, (1989)). This includes fragments such as acetonitrile, methanol, acetate, methyl ammonium, dimethyl ether, methane, and acetaldehyde.

Determination of the local energy minima in the binding site requires that many starting positions be sampled. This can be achieved by distributing, for example, 1,000-5,000 groups at random inside a sphere centered on the binding site; only the space not occupied by the protein needs to be considered. If the interaction energy of a particular group at a certain location with the protein is more positive than a given cut-off (*e.g.*, 5.0

kcal/mole) the group is discarded from that site. Given the set of starting positions, all the fragments are minimized simultaneously by use of the TDH approximation (Elber *et al.* (1990) *J. Am. Chem. Soc.* 112: 9161-9175). In this method, the forces on each fragment consist of its internal forces and those due to the protein. The essential element of this method is that the interactions between the fragments are omitted and the forces on the protein are normalized to those due to a single fragment. In this way simultaneous minimization or dynamics of any number of functional groups in the field of a single protein can be performed.

Minimization is performed successively on subsets of, *e.g.*, 100, of the randomly placed groups. After a certain number of step intervals, such as 1,000 intervals, the results can be examined to eliminate groups converging to the same minimum. This process is repeated until minimization is complete (*e.g.*, RMS gradient of 0.01 kcal/mole/Å). Thus the resulting energy minimized set of molecules comprises what amounts to a set of disconnected fragments in three dimensions representing potential side-chains for amino acid analogs.

The next step then is to connect the pieces with spacers assembled from small chemical entities (atoms, chains, or ring moieties) to form amino acid analogs, *e.g.*, each of the disconnected can be linked in space to generate a single molecule using such computer programs as, for example, NEWLEAD (Tschinke *et al.* (1993) *J. Med. Chem.* 36: 3863,3870). The procedure adopted by NEWLEAD executes the following sequence of commands (1) connect two isolated moieties, (2) retain the intermediate solutions for further processing, (3) repeat the above steps for each of the intermediate solutions until no disconnected units are found, and (4) output the final solutions, each of which is single molecule. Such a program can use for example, three types of spacers: library spacers, single-atom spacers, and fuse-ring spacers. The library spacers are optimized structures of small molecules such as ethylene, benzene and methylamide. The output produced by programs such as NEWLEAD consist of a set of molecules containing the

original fragments now connected by spacers. The atoms belonging to the input fragments maintain their original orientations in space. The molecules are chemically plausible because of the simple makeup of the spacers and functional groups, and energetically acceptable because of the rejection of solutions with van-der Waals radii violations.

In addition, the order in which the steps of this method are performed is purely illustrative in nature. In fact, the steps can be performed in any order or in parallel, unless otherwise indicated by the present disclosure.

Furthermore, the methods disclosed herein may be performed in either hardware, software, or any combination thereof, as those terms are currently known in the art. In particular, the present method may be carried out by software, firmware, or microcode operating on a computer or computers of any type. Additionally, software may comprise computer instructions in any form (*e.g.*, source code, object code, interpreted code, etc.) stored in any computer-readable medium (*e.g.*, ROM, RAM, magnetic media, punched tape or card, compact disc (CD) in any form, DVD, etc.). Furthermore, such software may also be in the form of a computer data signal embodied in a carrier wave, such as that found within the well-known Web pages transferred among devices connected to the Internet. Accordingly, certain embodiments are not limited to any particular platform, unless specifically stated otherwise in the present disclosure.

Exemplary computer hardware means suitable for carrying out certain embodiments can be a Silicon Graphics Power Challenge server with 10 R10000 processors running in parallel. Suitable software development environment includes CERIUS2 by Biosym/Molecular Simulations (San Diego, CA), or other equivalents.

The computational method described above has been effectively used in modifying enzymes of the protein synthesis machinery (*e.g.*, AARS) to allow incorporation of unnatural amino acids. The same suite of computational tools can also be leveraged to design the final products (*e.g.*, monoclonal

antibodies or other therapeutics) in which the unnatural amino acids would be incorporated so as to enhance or modify their structural or functional properties.

While particular embodiments disclosed herein have been shown and described, it will be apparent to those skilled in the art that changes and modifications may be made without departing from the broader aspect and, therefore, the appended claims are to encompass within their scope all such changes and modifications as fall within the true spirit of this invention.

Adoption of AARS from Different Organisms

A second strategy for generating an external mutant tRNA, modified or external mutant RS, or modified tRNA/RS pair involves importing a tRNA and/or synthetase from another organism into the translation system of interest, such as *Escherichia coli.* In this particular example, the properties of the heterologous synthetase candidate include, *e.g.*, that it does not charge *Escherichia coli* tRNA reasonably well (preferably not at all), and the properties of the heterologous tRNA candidate include, *e.g.*, that it is not acylated by *Escherichia coli* synthetase to a reasonable extent (preferably not at all).

Schimmel *et al.* reported that *Escherichia coli* GlnRS (EcGlnRS) does not acylate *Saccharomyces cerevisiae* tRNAGln (EcGlnRS lacks an N-terminal RNA-binding domain possessed by *Saccharomyces cerevisiae* GlnRS (ScGlnRS)). See, E. F. Whelihan and P. Schimmel, *EMBO J.*, 16:2968 (1997). For example, the *Saccharomyces cerevisiae* amber suppressor tRNAGln (SctRNAGlnCUA) was analyzed to determine whether it is also not a substrate for EcGlnRS. *In vitro* aminoacylation assays showed this to be the case; and *in vitro* suppression studies show that the SctRNAGlnCUA is competent in translation. See, *e.g.*, Liu and Schultz, *PNAS. U S A*, 96:4780 (1999). It was further shown that ScGlnRS does not acylate any *Escherichia coli* tRNA, only the SctRNAGlnCUA *in vitro.* The degree to which ScGlnRS is able to aminoacylate the SctRNAGlnCUA in *Escherichia coli* was also evaluated using an *in vivo* complementation assay. An amber nonsense mutation was

introduced at a permissive site in the β-lactamase gene. Suppression of the mutation by an amber suppressor tRNA should produce full-length β-lactamase and confer ampicillin resistance to the cell. When only SctRNAGlnCUA is expressed, cells exhibit an IC_{50} of 20 μg/mL ampicillin, indicating virtually no acylation by endogenous *Escherichia coli* synthetases; when SctRNAGlnCUA is coexpressed with ScGlnRS, cells acquire an IC_{50} of about 500 μg/mL ampicillin, demonstrating that ScGlnRS acylates SctRNAGlnCUA efficiently in *Escherichia coli*. See, Liu and Schultz, *PNAS, U S A*, 96:4780 (1999).

As another example, *Saccharomyces cerevisiae* $tRNA^{Asp}$ is known to be an external mutant to *Escherichia coli* synthetases. See, *e.g.*, Doctor and Mudd, *J. Biol. Chem.*, 238:3677 (1963); and, Kwok and Wong, *Can. J. Biochem.*, 58:213 (1980). It was demonstrated that an amber suppressor tRNA derived from it ($SctRNA^{Asp}_{CUA}$) is also an external mutant in *Escherichia coli* using the *in vivo* β-lactamase assay described above. However, the anticodon of $tRNA^{Asp}$ is a critical recognition element of AspRS, see, *e.g.*, Giege, *et al*, *Biochimie*, 78:605 (1996), and mutation of the anticodon to CUA results in a loss of affinity of the suppressor for AspRS. An *Escherichia coli* AspRS E93K mutant has been shown to recognize *Escherichia coli* amber suppressor $tRNA^{Asp}_{CUA}$ about an order of magnitude better than wt AspRS. See, *e.g.*, Martin, 'Thesis', Universite Louis Pasteur, Strasbourg, France, 1995. It was speculated that introduction of the related mutation in *Saccharomyces cerevisiae* AspRS (E188K) might restore its affinity for $SctRNA^{Asp}_{CUA}$. It was determined that the *Saccharomyces cerevisiae* AspRS(E188K) mutant does not acylate *Escherichia coli* tRNAs, but charges $SctRNA^{Asp}_{CUA}$ with moderate efficiency as shown by *in vitro* aminoacylation experiments. See, *e.g.*, Pastrnak, et al., *Helv. Chim. Acta*, 83:2277 (2000).

A similar approach involves the use of a heterologous synthetase as the external mutant synthetase and a mutant initiator tRNA of the same organism or a related organism as the modified tRNA. RajBhandary and coworkers found that an amber mutant of human initiator $tRNA^{fMet}$ is acylated

by *Escherichia coli* GlnRS and acts as an amber suppressor in yeast cells only when EcGlnRS is coexpressed. See, Kowal, et al., *PNAS U S A*, 98:2268 (2001). This pair thus represents an external mutant pair for use in yeast. Also, an *Escherichia coli* initiator $tRNA^{fMet}$ amber mutant was found that is inactive toward any *Escherichia coli* synthetases. A mutant yeast TyrRS was selected that charges this mutant tRNA, resulting in an external mutant pair in *Escherichia coli*.

Using the methods disclosed herein, the pairs and components of pairs desired above are evolved to generate external mutant tRNA and/or RS that possess desired characteristic, *e.g.*, that can preferentially aminoacylate an O-tRNA with an unnatural amino acid.

In certain embodiments, the modified tRNA and the modified RS can be derived by mutation of a naturally occurring tRNA and RS from a variety of organisms. In one embodiment, the modified tRNA and/or modified RS are derived from at least one organism, where the organism is a prokaryotic organism, *e.g.*, *Methanococcus jannaschii*, *Methanobacterium thennoautotrophicum*, *Halobacterium*, *Escherichia coli*, *A. fulgidus*, *P. furiosus*, *P. horikoshii*, *A. pernix*, *T. thermophilus*, or the like. Optionally, the organism is a eukaryotic organism, *e.g.*, plants (*e.g.*, complex plants such as monocots, or dicots), algea, fungi (*e.g.*, yeast, etc), animals (*e.g.*, mammals, insects, arthropods, etc.), insects, protists, or the like. Optionally, the modified tRNA is derived by mutation of a naturally occurring tRNA from a first organism and the modified RS is derived by mutation of a naturally occurring RS from a second organism. In one embodiment, the modified tRNA and modified RS can be derived from a mutated tRNA and mutated RS. In certain embodiments, the modified RS and/or modified tRNA from a first organism is provided to a translational system of a second organism, which optionally has non-functional endogenous RS and/or tRNA with respect to the codons recognized by the modified tRNA or modified RS.

The external mutant tRNA and/or the external mutant tRNA synthetase also can optionally be isolated from a variety of organisms. In one embodiment, the external mutant tRNA and/or external mutant synthetase are isolated from at least one organism, where the organism is a prokaryotic organism, *e.g.*, *Methanococcus jannaschii, Methanobacterium thermoautotrophicum*, Halobacterium, *Escherichia coli*, A. *fulgidus, P. furiosus*, P. *horikoshii*, A. *pernix*, T. *thermophilus*, or the like. Optionally, the organism is a eukaryotic organism, *e.g.*, plants (*e.g.*, complex plants such as monocots, or dicots), algea, fungi (*e.g.*, yeast, etc), animals (*e.g.*, mammals, insects, arthropods, etc.), insects, protists, or the like. Optionally, the external tRNA is isolated from a naturally occurring tRNA from a first organism and the external mutant synthetase is isolated from a naturally occurring RS from a second organism. In one embodiment, the external mutant tRNA and/or external mutant tRNA synthetase can be isolated from one or more library (which optionally comprises one or more tRNA and/or RS from one or more organism (including those comprising prokaryotes and/or eukaryotes).)

Methods for selecting an external mutant tRNA and/or tRNA synthetase pair for use in any translation system are also disclosed herein. The methods include: introducing a marker gene, a tRNA and/or an aminoacyl-tRNA synthetase (RS) isolated or derived from a first organism into a first set of cells from the second organism; introducing the marker gene and the tRNA or RS into a duplicate cell set from the second organism; and, selecting for surviving cells in the first set that fail to survive in the duplicate cell set, where the first set and the duplicate cell set are grown in the presence of a selection agent, and where the surviving cells comprise the external mutant tRNAand/or RS for use in the in a translation system. In one embodiment, comparing and selecting includes an *in vivo* complementation assay. In another embodiment, the concentration of the selection agent is varied. The same assay may also be conducted in an *in vitro* or *in vivo* system based on the second organism.

Generation of AARS by Mutagenesis and Selection / Screening

The mutation or modification of an AARS to be used for incorporation of a non-natural amino acid into a target polypeptide or protein can be performed by using directed mutagenesis once the desired contact amino acid residues have been identified. Identification of the contact amino acids can be performed using any method that allows analysis of the structure of the AARS, including crystallographic analysis, computer modeling, nuclear magnetic resonance (NMR) spectroscopy, library screening, or a combination of any of these or other methods.

A number of AARS molecules have been sequenced, and provide guidance as to which amino acids are important for binding the amino acid with which to charge the corresponding tRNA. See, for example, SEQ ID Nos. 48-103.

In certain embodiments, the AARS capable of charging a particular external mutant tRNA with a particular unnatural amino acid can be obtained by mutagenesis of the AARS to generate a library of candidates, followed by screening and/or selection of the candidate AARS's capable of their desired function. Such external mutant AARSs and external mutant tRNAs may be used for *in vitro / in vivo* production of desired proteins with modified unnatural amino acids.

Thus methods for generating components of the protein biosynthetic machinery, such as the external mutant RSs, external mutant tRNAs, and/or external mutant tRNA/RS pairs that can be used to incorporate an unnatural amino acid are provided in certain embodiments disclosed herein.

In one embodiment, methods for producing at least one recombinant external mutant aminoacyl-tRNA synthetase comprise: (a) generating a library of (optionally mutant) RSs derived from at least one aminoacyl-tRNA synthetase (RS) from a first organism, *e.g.*, a eukaryotic organism (such as a yeast), or a prokaryotic organism, such as *Methanococcus jannaschii*, *Methanobacterium thennoautotrophicum*, *Halobacterium*,

Escherichia coli, A. fulgidus, P. furiosus, P. horikoshii, A. pernix, T. thermophilus, or the like; (b) selecting (and/or screening) the library of RSs (optionally mutant RSs) for members that aminoacylate an external mutant tRNA in the presence of an unnatural amino acid and a natural amino acid, thereby providing a pool of active (optionally mutant) RSs; and/or, (c) selecting (optionally through negative selection) the pool for active RSs (*e.g.*, mutant RSs) that preferentially aminoacylate the O-tRNA in the absence of the unnatural amino acid, thereby providing the at least one recombinant external mutant synthetase, wherein the at least one recombinant external mutant synthetase preferentially aminoacylates the external mutant tRNA with the unnatural amino acid. Recombinant external mutant synthetases produced by the methods are also included in certain embodiments disclosed herein.

In one embodiment, the RS is an inactive RS, which may have been generated from mutating an active RS. For example, the inactive RS can be generated by mutating at least about 1, at least about 2, at least about 3, at least about 4, at least about 5, at least about 6, or at least about 10 or more amino acids to different amino acids, *e.g.*, alanine.

Libraries of mutant RSs can be generated using various mutagenesis techniques known in the art. For example, the mutant RSs can be generated by site-specific mutations, random mutations, diversity generating recombination mutations, chimeric constructs, and by other methods described herein or known in the art.

In one embodiment, selecting (and/or screening) the library of RSs (optionaly mutant RSs) for members that are active, *e.g.*, that aminoacylate an external mutant tRNA in the presence of an unnatural amino acid and a natural amino acid, includes: introducing a positive selection or screening marker, *e.g.*, an antibiotic resistance gene, or the like, and the library of (optionally mutant) RSs into a plurality of cells, wherein the positive selection and/or screening marker comprises at least one codon, whose translation (optionally conditionally) depends on the ability of a candidate RSs to charge

the external mutant tRNA (with either a natural and/or a unnatural amino acid); growing the plurality of cells in the presence of a selection agent; identifying cells that survive (or show a specific response) in the presence of the selection and/or screening agent by successfully translate the codon in the positive selection or screening marker, thereby providing a subset of positively selected cells that contains the pool of active (optionally mutant) RSs. Optionally, the selection and/or screening agent concentration can be varied. In certain embodiments, the cells do not contain a functional endogenous tRNA, RS or tRNA-RS pair that can help to translate the codon. The endogenous tRNA / RS pair may be disabled by gene deletion and/or RS inhibitors.

Since many essential genes of the cell likely also contain such codon that depends on the ability of the external mutant synthetase to charge the modified tRNA at the absence of functional endogenous RS / tRNA pair, in one embodiment, no extra positive selection markers are needed for the positive selection process – the survival of the cell can be used as a readout of the positive selection process.

In one aspect, the positive selection marker is a chloramphenicol acetyltransferase (CAT) gene. Optionally, the positive selection marker is a β-lactamase gene. In another aspect the positive screening marker comprises a fluorescent or luminescent screening marker or an affinity based screening marker (*e.g.*, a cell surface marker).

In a similar embodiment, a cell-free *in vitro* system may be used to test the ability of the external mutant synthetase to charge the modified tRNA in a positive screening. For example, the ability of the *in vitro* system to translate a positive screening gene, such as a fluorescent marker gene, may depend on the ability of the external mutant synthetase to charge modified tRNA to read through a codon of the marker gene.

In one embodiment, negatively selecting or screening the pool for active RSs (optionally mutants) that preferentially aminoacylate the mutant tRNA in the absence of the unnatural amino acid includes: introducing a

negative selection or screening marker with the pool of active (optionally mutant) RSs from the positive selection or screening into a plurality of translational system, wherein the negative selection or screening marker comprises at least one codon (*e.g.*, codon for a toxic marker gene, *e.g.*, a ribonuclease barnase gene), whose translation depends on the ability of a candidate RS to charge the external mutant tRNA (with a natural amino acid); and, identifying the translation system that shows a specific screening response in a first media supplemented with the unnatural amino acid and a screening or selection agent, but fail to show the specific response in a second media supplemented with the natural amino acid and the selection or screening agent, thereby providing surviving cells or screened cells with the at least one recombinant RS.

In one aspect, the concentration of the selection (and/or screening) agent is varied. In some aspects the first and second organisms are different. Thus, the first and/or second organism optionally comprises: a prokaryote, a eukaryote, a mammal, an *Escherichia coli,* a fungi, a yeast, an archaebacterium, a eubacterium, a plant, an insect, a protist, etc. In other embodiments, the screening marker comprises a fluorescent or luminescent screening marker (such as green fluorescent protein) or an affinity based screening marker.

Also, some aspects include wherein the negative selction marker comprises a ribonuclease barnase gene (which comprises at least one said codon). Other aspects include wherein the screening marker optionally comprises a fluorescent or luminescent screening marker or an affinity based screening marker. In the embodiments herein, the screenings and/or selections optionally include variation of the screening and/or selection stringency.

In one aspect, the second set of mutated RS derived from at least one recombinant RS can be generated by mutagenesis, *e.g.*, random mutagenesis, site-specific mutagenesis, recombination or a combination thereof.

The methods embodied herein optionally comprise wherein the unnatural amino acid is selected from, *e.g.*: an O-methyl-L-tyrosine, an L-3-(2-naphthyl)alanine, a 3-methyl-phenylalanine, an O-4-allyl-L-tyrosine, a 4-propyl-L-tyrosine, a tri-O-acetyl-GlcNAcβ-serine, an L-Dopa, a fluorinated phenylalanine, an isopropyl-L-phenylalanine, a p-azido-L-phenylalanine, a p-acyl-L-phenylalanine, a p-benzoyl-L-phenylalanine, an L-phosphoserine, a phosphonoserine, a phosphonotyrosine, a p-iodo-phenylalanine, a p-bromophenylalanine, a p-amino-L-phenylalanine, and an isopropyl-L-phenylalanine. A recombinant RS produced by the methods herein is also included in the embodiments disclosed herein.

In a related aspect, methods for producing a recombinant external mutant tRNA include: (a) generating a library of mutant tRNAs derived from at least one tRNA, from a first organism; (b) selecting (*e.g.*, negatively selecting) or screening the library for (optionally mutant) tRNAs that are aminoacylated by an aminoacyl-tRNA synthetase (RS) from a second organism in the absence of a RS from the first organism, thereby providing a pool of tRNAs (optionally mutant); and, (c) selecting or screening the pool of tRNAs (optionally mutant) for members that are aminoacylated by an introduced external mutant RS, thereby providing at least one recombinant tRNA; wherein the at least one recombinant tRNA recognizes a non-natural amino acid codon and is not efficiency recognized by the RS from the second organism and is preferentially aminoacylated by the external mutant RS.

The various methods disclosed herein ptionally comprise wherein selecting or screening comprises one or more positive or negative selection or screening, *e.g.*, a change in amino acid permeability, a change in translation efficiency, and a change in translational fidelity. Additionally, the one or more change is optionally based upon a mutation in one or more gene in an organism in which an external mutant tRNA-tRNA synthetase pair are used to produce such protein. Selecting and/or screening herein optionally comprises wherein at least 2 codons within one or more selection gene or within one or more

screening gene are used. Such multiple codons are optionally within the same gene or within different screening/selection genes. Additionally, the optional multiple codons are optionally different codons or comprise the same type of codons.

Kits are an additional feature of certain embodiments disclosed herein. For example, the kits can include one or more translation system as noted above (*e.g.*, a cell), one or more tRNA (including modified or mutated tRNA), one or more AARS (including modified or mutated AARS), one or more unnatural amino acid, *e.g.*, with appropriate packaging material, containers for holding the components of the kit, instructional materials for practicing the methods herein and/or the like. If one or more AARS and/or one or more tRNA are provided in a kit, they may be supplied as nucleic acids, or proteins and may be part of a single vector or contained in separate vectors. Similarly, products of the translation systems (*e.g.*, proteins such as EPO analogues comprising unnatural amino acids) can be provided in kit form, *e.g.*, with containers for holding the components of the kit, instructional materials for practicing the methods herein and/or the like.

Exemplary Uses

Well over 100 non-coded amino acids (all ribosomally acceptable) have been reportedly introduced into proteins using other methods (see, for example, Schultz *et al.*, *J. Am. Chem. Soc.*, 103: 1563-1567, 1981; Hinsberg *et al.*, *J. Am. Chem. Soc.*, 104: 766-773, 1982; Pollack *et al.*, *Science*, 242: 1038-1040, 1988; Nowak *et al.*, *Science*, 268: 439-442, 1995) all these analogs may be used in the subject methods for efficient incorporation of these analogs into protein products.

In another preferred embodiment, two or more analogs may be used in the same *in vitro* or *in vivo* translation system, each with its external mutant tRNA or external mutant synthetase pairs. This is more easily accomplished when a natural amino acid is encoded by four or more codons

(such as six for Leu and Arg). However, for amino acids encoded by only two codons, one can be reserved for the natural amino acid, while the other "shared" by one or more amino acid analog(s). These analogs may resemble only one natural amino acid (for example, different Phe analogs), or resemble different amino acids (for example, analogs of Phe and Tyr).

In certain embodiments, a first nucleic acid encoding an external mutant / modified tRNA molecule that is not charged efficiently by an endogenous aminoacyl-tRNA synthetase in the cell / *in vitro* translation system (IVT), or the external mutant / modified tRNA itself. According to some embodiments, a second nucleic acid encoding an external mutant / modified aminoacyl tRNA synthetase (AARS) is also introduced into the cell / IVT. The external mutant / modified AARS is capable of charging the external mutant / modified tRNA with a chosen amino acid analog. The amino acid analog can then be provided to the cell so that it can be incorporated into one or more proteins within the cell or IVT.

In other embodiments, the environment is a cell. A variety of cells (or lysates thereof suitable for IVT) can be used in certain methods, including, for example, a bacterial cell, a fungal cell, an insect cell, and a mammalian cell (*e.g.*, a human cell or a non-human mammal cell). In one embodiment, the cell is an *E. coli* cell, in another embodiment, the cell is a *Pseudomonas* cell.

In certain embodiments, the amino acid analog can be provided by directly contacting the cell or IVT with the analog, for example, by applying a solution of the analog to the cell in culture, or by directly adding the analog to the IVT. The analog can also be provided by introducing one or more additional nucleic acid construct(s) into the cell / IVT, wherein the additional nucleic acid construct(s) encodes one or more amino acid analog synthesis proteins that are necessary for synthesis of the desired analog.

Certain embodiments further involve introducing a template nucleic acid construct into the cell / IVT, the template encoding a protein, wherein the nucleic acid construct contains at least one degenerate codon

sequence. The nucleic acids introduced into the cell / IVT can be introduced as one construct or as a plurality of constructs. In certain embodiments, the various nucleic acids are included in the same construct. For example, the nucleic acids can be introduced in any suitable vectors capable of expressing the encoded tRNA and/or proteins in the cell / IVT. In one embodiment, the first and second nucleic acid sequences are provided in one or more plasmids. In another embodiment, the vector or vectors used are viral vectors, including, for example, adenoviral and lentiviral vectors. The sequences can be introduced with an appropriate promoter sequence for the cell / IVT, or multiple sequences that can be inducible for controlling the expression of the sequences.

For *in vitro* use, one or more external mutant synthetase can be recombinantly produced and supplied to any the available *in vitro* translation systems (such as the commercially available Wheat Germ Lysate-based PROTEINscript-PRO™, Ambion®'s *E. coli* system for coupled *in vitro* transcription/translation; or the rabbit reticulocyte lysate-based Retic Lysate IVT™ Kit from Ambion®). Optionally, the *in vitro* translation system can be selectively depleted of one or more natural AARSs (by, for example, immunodepletion using immobilized antibodies against natural AARS) and/or natural amino acids so that enhanced incorporation of the analog can be achieved. Alternatively, nucleic acids encoding the re-designed external mutant synthetases may be supplied in place of recombinantly produced AARSs. The *in vitro* translation system is also supplied with the analogs to be incorporated into mature protein products.

Although *in vitro* protein synthesis usually cannot be carried out on the same scale as *in vivo* synthesis, *in vitro* methods can yield hundreds of micrograms of purified protein containing amino acid analogs. Such proteins have been produced in quantities sufficient for their characterization using circular dichroism (CD), nuclear magnetic resonance (NMR) spectrometry, and X-ray crystallography. This methodology can also be used to investigate the role of hydrophobicity, packing, side chain entropy and hydrogen bonding in

determining protein stability and folding. It can also be used to probe catalytic mechanism, signal transduction and electron transfer in proteins. In addition, the properties of proteins can be modified using this methodology. For example, photocaged proteins can be generated that can be activated by photolysis, and novel chemical handles have been introduced into proteins for the site specific incorporation of optical and other spectroscopic probes.

The development of a general approach for the incorporation of amino acid analogs into proteins *in vivo*, directly from the growth media, would greatly enhance the power of unnatural amino acid mutagenesis. For example, the ability to synthesize large quantities of proteins containing heavy atoms would facilitate protein structure determination, and the ability to site-specifically substitute fluorophores or photocleavable groups into proteins in living cells would provide powerful tools for studying protein function *in vivo*. Alternatively, one might be able to enhance the properties of proteins by providing building blocks with new functional groups, such as a keto-containing amino acid.

For *in vivo* use, one or more AARS can be supplied to a host cell (prokaryotic or eukaryotic) as genetic materials, such as coding sequences on plasmids or viral vectors, which may optionally integrate into the host genome and constitutively or inducibly express the re-designed AARSs. A heterologous or endogenous protein of interest can be expressed in such a host cell, at the presence of supplied amino acid analogs. The protein products can then be purified using any art-recognized protein purification techniques, or techniques specially designed for the protein of interest.

These are a few possible means for generating a transcript which encodes a polypeptide. In general, any means known in the art for generating transcripts can be employed to synthesize proteins with amino acid analogs. For example, any *in vitro* transcription system or coupled transcription / translation systems can be used for generate a transcript of interest, which then serves as a template for protein synthesis. Alternatively, any cell, engineered cell / cell line, or functional components (lysates, membrane fractions, etc.) that

is capable of expressing proteins from genetic materials can be used to generate a transcript. These means for generating a transcript will typically include such components as RNA polymerase (T7, SP6, etc.) and co-factors, nucleotides (ATP, CTP, GTP, UTP), necessary transcription factors, and appropriate buffer conditions, as well as at least one suitable DNA template, but other components may also added for optimized reaction condition. A skilled artisan would readily envision other embodiments similar to those described herein.

Chemical Moieties

In certain embodiments, the unnatural amino acid(s) and/or the therapeutic molecule comprises a chemically reactive moiety. The moiety may be strongly electrophilic or nucleophilic and thereby be available for reacting directly with the therapeutic molecule or the antibody or fragment thereof. Alternatively, the moiety may be a weaker electrophile or nucleophile and therefore require activation prior to the conjugation with the therapeutic molecule or the antibody or fragment thereof. This alternative would be desirable where it is necessary to delay activation of the chemically reactive moiety until an agent is added to the molecule in order to prevent the reaction of the agent with the moiety. In either scenario, the moiety is chemically reactive, the scenarios differ (in the reacting with antibody scenario) by whether following addition of an agent, the moiety is reacted directly with an antibody or fragment thereof or is reacted first with one or more chemicals to render the moiety capable of reacting with an antibody or fragment thereof. In certain embodiments, the chemically reactive moiety includes an amino group, a sulfhydryl group, a hydroxyl group, a carbonyl-containing group, or an alkyl leaving group.

Certain embodiments may employ click chemistry, which include, but is not limited to, Huisgen 1, 3-dipolar cycloaddition, in particular the Cu(I)-catalyzed stepwise variant, Diels-Alder reaction, nucleophilic substitution

especially to small strained rings like epoxy and aziridine compounds, carbonyl-chemistry-like formation of ureas and amides, addition reactions to carbon-carbon double bonds like epoxidation and dihydroxylation.

Thus, in addition to or instead of glycosylation of polypeptides of the embodiments disclosed herein, other chemical moieties (including poly(ethylene) glycol) may be added, linked, joined, or otherwise conjugated or incorporated into the modified polypeptides. PEGylation is a process to covalently attach oligosaccharides and synthetic polymers such as polyethylene glycol (PEG) site-specifically onto therapeutic protein molecules. PEGylation can significantly enhance protein half-life by shielding the polypeptide from proteolytic enzymes and increasing the apparent size of the protein, thus reducing clearance rates. Moreover, PEG conjugates can enhance protein solubility and have beneficial effects on biodistribution. The physical and pharmacological properties of PEGylated proteins are affected by the number and the size of PEG chains attached to the polypeptide, the location of the PEG sites, and the chemistry used for PEGylation.

Examples of PEG conjugation to proteins include reactions of N-hydroxysuccinimidyl ester derivatized PEGs with lysine, 1,4-addition reactions of maleimide and vinylsulfone derivatized PEGs with cysteine, and condensation of hydrazide containing PEGs with aldehydes generated by oxidation of glycoproteins. When more than one reactive site is present in a protein (*e.g.*, multiple amino or thiol groups) or reactive electrophiles are used, nonselective attachment of one or multiple PEG molecules can occur, leading to the generation of a heterogeneous mixture that is difficult to separate. The lack of selectivity and positional control in the attachment of PEG chains can lead to significant losses in biological activity and possibly enhanced immunogenicity of the conjugated protein. In fact, historically, loss of biological activity and product heterogeneity have been the two most common problems encountered in the development of long-acting protein pharmaceuticals using standard PEGylation techniques. Modification of proteins with amine-reactive

PEGs typically results in drastic loss of biological activity due to modification of lysine residues located in regions of the protein important for biological activity. In certain situations, bioactivity of growth hormones may be reduced 400-fold or more. For example, bioactivity of GCSF is reduced 1,000-fold when the proteins are modified using conventional amine-PEGylation technologies (Clark *et al.*, *J. Biol. Chem.* 271: 21969, 1996; Bowen *et al.*, *Exp. Hematol.* 27, 425, 1999). Thus there is a need for a method that allows for the completely site-specific and irreversible attachment of PEG chains to proteins.

It would be advantageous to use advanced protein engineering technologies to create long-acting, "patient friendly" human protein pharmaceuticals, by, for example, incorporating unnatural amino acids into a drug protein, such that the engineered drug protein may achieve longer half life and/or sustained or even enhanced biological activity. Towards this end, certain embodiments disclosed herein may be used to overcome problems such as heterogeneity and loss of activity inherent in standard amine-PEGylation techniques. Incorporating unnatural amino acids will provide unique, pre-determined sites away from the binding or the catalytic site on the target protein where PEG molecules can be site-specifically conjugated. In addition, PEG molecules may be attached to unnatural amino acids through techniques other than amine-PEGylation, thus sparing the primary amine groups of lysines from undesirable PEGylation. These techniques may be used to enhance the half-life, efficacy, and/or safety of bio-pharmaceuticals in all areas, including the specific field of cancer, endocrinology, infectious disease, and inflammation, etc.

As an illustrative example, Click Chemistry or cycloaddition may be used to form a triazole linkage. One particular example of cycloaddition is a copper-mediated Huisgen [3+2] cycloaddition (Tornoe *et al.*, *J. Org. Chem.* 67: 3057, 2002; Rostovtsev *et al.*, *Angew. Chem., Int. Ed.* 41: 596, 2002; and Wang *et al.*, *J. Am. Chem. Soc.* 125: 3192, 2003) of an azide and an alkyne is external mutant to all functional groups found in proteins, and forms a stable

triazole linkage, this reaction can be used for the selective PEGylation of proteins. For example, Deiters *et al.* (*Bioorg. Med. Chem. Lett.* 14(23): 5743-5745, 2004) report a generally applicable PEGylation methodology based on the site-specific incorporation of para-azidophenylalanine into proteins in yeast. The azido group was used in a mild [3+2] cycloaddition reaction with an alkyne derivatized PEG reagent to afford selectively PEGylated protein. This strategy should be useful for the generation of selectively PEGylated proteins for therapeutic applications.

In certain embodiments, the polypeptide is a therapeutic, diagnostic, or other protein selected from: Alpha-1 antitrypsin, Angiostatin, Antihemolytic factor, antibodies (including an antibody or a functional fragment or derivative thereof selected from: Fab, Fab', F(ab)2, Fd, Fv, ScFv, diabody, tribody, tetrabody, dimmer, trimer or minibody), angiogenic molecules, angiostatic molecules, Apolipoprotein, Apoprotein, Atrial natriuretic factor, Atrial natriuretic polypeptide, Asparaginase, Adenosine deaminase, Hirudin, Ciliary Neurotrophic factor, bone morphogenic factor (any and all BMPs), Atrial peptides, C-X-C chemokines (*e.g.*, T39765, NAP-2, ENA-78, Gro-a, Gro-b, Gro-c, IP-10, GCP-2, NAP-4, SDF-1, PF4, MIG), Calcitonin, CC chemokines (*e.g.*, Monocyte chemoattractant protein-1, Monocyte chemoattractant protein-2, Monocyte chemoattractant protein-3, Monocyte inflammatory protein-1 alpha, Monocyte inflammatory protein-1 beta, RANTES, I309, R83915, R91733, HCC1, T58847, D31065, T64262), CD40 ligand, calcitonin, C-kit ligand, collagen, Colony stimulating factor (CSF), C-type natriuretic peptide (CNP), Complement factor 5a, Complement inhibitor, Complement receptor 1, cytokines, (*e.g.*, epithelial Neutrophil Activating Peptide-78, GROα/MGSA, GROβ, GROγ, MIP-1α, MIP-1δ, MCP-1), deoxyribonucleic acids, Epidermal Growth Factor (EGF), Erythropoietin, Exfoliating toxins A and B, Factor IX, Factor VII, Factor VIII, Factor X, Fibroblast Growth Factor (FGF), Fibrinogen, Fibronectin, granulocyte- colony stimulating factor (G-CSF), granulocyte macrophage colony stimulating factor (GM-CSF), follitropin,

Glucocerebrosidase, Gonadotropin, glucagons, GLP-1, growth factors, Hedgehog proteins (*e.g.*, Sonic, Indian, Desert), Human Growth Hormone, Hemoglobin, Hepatocyte Growth Factor (HGF), Hepatitis viruses, Hirudin, Human serum albumin, Insulin, Insulin-like Growth Factor (IGF), interferons (*e.g.*, IFN-α, IFN-β, IFN-γ, IFN-ε, IFN-ζ, IFN-η, IFN-κ, IFN-λ, IFN-τ, IFN-ς, IFN-ω), interleukins (*e.g.*, IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, IL-15, etc.), Keratinocyte Growth Factor (KGF), Lactoferrin, leukemia inhibitory factor, Luciferase, Luteinizing hormone, Neurturin, Neutrophil inhibitory factor (NIF), oncostatin M, Osteogenic protein, Parathyroid hormone, PD-ECSF, PDGF, peptide hormones (*e.g.*, Human Growth Hormone), Pleiotropin, Protein A, Protein G, Phenylalanine hydroxylase, Parathormone (PTH), Prolactin, Pyrogenic exotoxins A, B, and C, Relaxin, Renin, ribonucleic acids, SCF, Soluble complement receptor I, Soluble I-CAM 1, Soluble interleukin receptors (IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, 1IL-5), Soluble TNF receptor, Somatomedin, Somatostatin, Somatotropin, Streptokinase, Superantigens, *i.e.*, Staphylococcal enterotoxins (SEA, SEB, SEC1, SEC2, SEC3, SED, SEE), Superoxide dismutase (SOD), Toxic shock syndrome toxin (TSST-1), Thymosin alpha 1, Tissue plasminogen activator, Tumor necrosis factor beta (TNF beta), Tumor necrosis factor receptor (TNFR), Tumor necrosis factor-alpha (TNF alpha), Tumor necrosis factor related apoptosis-inducing ligand (TRAIL), Vascular Endothelial Growth Factor (VEGEF), Urokinase; a transcriptional modulator that modulates cell growth, differentiation, or regulation, wherein the transcriptional modulator is from prokaryotes, viruses, or eukaryotes, including fungi, plants, yeasts, insects, and animals, including mammals; expression activator selected from cytokines, inflammatory molecules, growth factors, their receptors, oncogene products, interleukins (*e.g.*, IL-1, IL-2, IL-8, etc.), interferons, FGF, IGF-I, IGF-II, FGF, PDGF, TNF, TGF-α, TGF-β, EGF, KGF, SCF/c-Kit, CD40L/CD40, VLA-4/VCAM-1, ICAM-1/LFA-1, and hyalurin/CD44; signal transduction molecules and corresponding oncogene products, *e.g.*, Mos, Ras,

Raf, and Met; transcriptional activators and suppressors, *e.g.*, p53, Tat, Fos, Myc, Jun, Myb, Rel; steroid hormone receptors selected from receptors for estrogen, progesterone, testosterone, aldosterone, LDL, or corticosterone; or an enzyme selected from: amidases, amino acid racemases, acylases, dehalogenases, dioxygenases, diarylpropane peroxidases, epimerases, epoxide hydrolases, esterases, isomerases, kinases, glucose isomerases, glycosidases, glycosyl transferases, haloperoxidases, monooxygenases (*e.g.*, p450s), lipases, lignin peroxidases, nitrile hydratases, nitrilases, proteases, phosphatases, subtilisins, transaminase, or nucleases.

In the event that the protein or molecule of interest to be modified is an antibody or antibody fragment, the non-natural amino acid residue(s) may be placed at any location or position in the antibody structure, depending on the desired goal. For example, the non-natural amino acid residue may be placed in the Fab variable region, the Fc region, or in another location that interacts with the Fc region of the antibody. In other embodiments, the non-natural amino acid residue may be placed in the binding interface of the antibody, or the V_H region. In certain embodiments, the modified antibody exhibits an increase or decrease in its ability to kill one or more targets. In particular, an antibody with increased ability to kill one or more targets, or with reduced side effects may be desired.

In other embodiments, the non-natural amino acid(s) confer enhanced binding affinity to an Fc-receptor and/or to C1q of the complement system. In particular, a modified antibody may have an altered (*e.g.*, enhanced) affinity and/or specificity for an antigen or a protein binding partner (*e.g.*, C1q of the complement and/or the Fc receptor on macrophages, etc.). For example, modification of a molecule may increase or decrease its antibody-dependent cell-mediated cytotoxicty (ADCC) function, or complement fixation activity. In other examples, modification of a particular molecule may increase or decrease its ability to bind another molecule of natural counter structure (such as an antibody).

Glycosylation through Unnatural Amino Acids

The post-translational modification of proteins by glycosylation can affect protein folding and stability, modify the intrinsic activity of proteins, and modulate their interactions with other biomolecules. See, *e.g.*, Varki, *Glycobiology* 3: 97-130, 1993. Natural glycoproteins are often present as a population of many different glycoforms, which makes analysis of glycan structure and the study of glycosylation effects on protein structure and function difficult. Therefore, methods for the synthesis of natural and unnatural homogeneously glycosylated proteins are needed for the systematic understanding of glycan function, and for the development of improved glycoprotein therapeutics.

One previously known approach for making proteins having desired glycosylation patterns makes use of glycosidases to convert a heterogeneous natural glycoprotein to a simple homogenous core, onto which saccharides can then be grafted sequentially with glycosyl transferases. See, *e.g.*, Witte *et al.*, *J. Am. Chem. Soc.* 119: 2114-2118, 1997. A limitation of this approach is that the primary glycosylation sites are predetermined by the cell line in which the protein is expressed. Alternatively, a glycopeptide containing the desired glycan structure can be synthesized by solid phase peptide synthesis. This glycopeptide can be coupled to other peptides or recombinant protein fragments to afford a larger glycoprotein by native chemical ligation (see, *e.g.*, Shin *et al.*, *J. Am. Chem. Soc.* 121: 11684-11689, 1999), expressed protein ligation (see, *e.g.*, Tolbert and Wong, *J. Am. Chem. Soc.* 122: 5421-5428, 2000), or with engineered proteases (see, *e.g.*, Witte *et al.*, *J. Am. Chem. Soc.* 120: 1979-1989, 1998). Both native chemical ligation and expressed protein ligation are most effective with small proteins, and necessitate a cysteine residue at the N-terminus of the glycopeptide.

When a protease is used to ligate peptides together, the ligation site must be placed far away from the glycosylation site for good coupling yields. See, *e.g.*, Witte *et al.*, *J. Am. Chem. Soc.* 120: 1979-1989, 1998. A

third approach is to modify proteins with saccharides directly using chemical methods. Good selectivity can be achieved with haloacetamide saccharide derivatives, which are coupled to the thiol group of cysteine (see, *e.g.*, Davis and Flitsch, *Tetrahedron Lett.* 32: 6793-6796, 1991; and Macmillan *et al.*, *Org. Lett.* 4: 1467-1470, 2002). But this method can become problematic with proteins that have more than one cysteine residue.

Certain embodiments provided herein disclose methods for synthesis of glycoproteins. These methods involve, in some embodiments, incorporating into a protein an unnatural amino acid that comprises a first reactive group; and contacting the protein with a saccharide moiety that comprises a second reactive group, wherein the first reactive group reacts with the second reactive group, thereby forming a covalent bond that attaches the saccharide moiety to the unnatural amino acid of the protein. Glycoproteins produced by these methods are also included in certain embodiments.

The first reactive group is, in some embodiments, an electrophilic moiety (*e.g.*, a keto moiety, an aldehyde moiety, and/or the like), and the second reactive group is a nucleophilic moiety. In some embodiments, the first reactive group is a nucleophilic moiety and the second reactive group is an electrophilic moiety (*e.g.*, a keto moiety, an aldehyde moiety, and/or the like). For example, an electrophilic moiety is attached to the saccharide moiety and the nucleophilic moiety is attached to the unnatural amino acid. The saccharide moiety can include a single carbohydrate moiety, or the saccharide moiety can include two or more carbohydrate moieties.

In some embodiments, the methods further involve contacting the saccharide moiety with a glycosyl transferase, a sugar donor moiety, and other reactants required for glycosyl transferase activity for a sufficient time and under appropriate conditions to transfer a sugar from the sugar donor moiety to the saccharide moiety. The product of this reaction can, if desired, be contacted by at least a second glycosyl transferase, together with the appropriate sugar donor moiety.

In certain embodiments, the method further comprises contacting the saccharide moiety with one or more of a β1-4N-acetylglucosaminyl transferase, an α1,3-fucosyl transferase, an α1,2-fucosyl transferase, an α1,4-fucosyl transferase, a β1-4-galactosyl transferase, a sialyl transferase, and/or the like, to form a biantennary or triantennary oligosaccharide structure. In one embodiment, the saccharide moiety comprises a terminal GlcNAc, the sugar donor moiety is UDP-Gal and the glycosyl transferase is a β-1,4-galactosyl transferase.

In one embodiment, the saccharide moiety comprises a terminal GlcNAc, the sugar donor moiety is UDP-GlcNAc and the glycosyl transferase is a β1-4N-acetylglucosaminyl transferase.

Optionally, the some methods further comprise contacting the product of the N-acetylglucosaminyl transferase reaction with a β1-4mannosyl transferase and GDP-mannose to form a saccharide moiety that comprises Manβ1-4GlcNAcβ1-4GlcNAc-. Optionally, the method further comprises contacting the Manβ1-4GlcNAcβ1-4GlcNAc-moiety with an α1-3mannosyl transferase and GDP-mannose to form a saccharide moiety that comprises Manα1-3Manβ1-4GlcNAcβ1-4GlcNAc-. Optionally, the method further comprises contacting the Manα1-3Manβ1-4GlcNAcβ1-4GlcNAc- moiety with an α1-6 mannosyl transferase and GDP-mannose to form a saccharide moiety that comprises Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-. Optionally, the method further comprises contacting the Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-moiety with a β1-2N-acetylglucosaminyl transferase and UDP-GlcNAc to form a saccharide moiety that comprises Manα1-6(GlcNAcβ1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-. Optionally, the method further comprises contacting the Manα1-6(GlcNAcβ1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-moiety with a β1-2N-acetylglucosaminyl transferase and UDP-GlcNAc to form a saccharide moiety that comprises GlcNAcβ1-2Manα1-6(GlcNAcβ1-2Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc-.

The step of incorporating into a protein an unnatural amino acid that comprises a first reactive group, in some embodiments, comprises using an external mutant tRNA, an external mutant RS, or an external mutant tRNA/RS pair. In such cases, the external mutant tRNA preferentially recognizes a degenerate codon for wild-type tRNA, and incorporates the unnatural amino acid into the protein in response to the degenerate codon, and wherein the external mutant synthetase preferentially aminoacylates the external mutant tRNA with the unnatural amino acid. In some embodiments, the unnatural amino acid is incorporated into the polypeptide *in vivo*.

A wide variety of suitable reactive groups are known to those of skill in the art. Such suitable reactive groups can include, for example, amino, hydroxyl, carboxyl, carboxylate, carbonyl, alkenyl, alkynyl, aldehyde, ester, ether (*e.g.*, thio-ether), amide, amine, nitrile, vinyl, sulfide, sulfonyl, phosphoryl, or similarly chemically reactive groups. Additional suitable reactive groups include, but are not limited to, maleimide, N hydroxysuccinimide, sulfo-N-hydroxysuccinimide, nitrilotriacetic acid, activated hydroxyl, haloacetyl (*e.g.*, bromoacetyl, iodoacetyl), activated carboxyl, hydrazide, epoxy, aziridine, sulfonylchloride, trifluoromethyldiaziridine, pyridyldisulfide, N-acyl-imidazole, imidazolecarbamate, vinylsulfone, succinimidylcarbonate, arylazide, anhydride, diazoacetate, benzophenone, isothiocyanate, isocyanate, imidoester, fluorobenzene, biotin and avidin.

In some embodiments, one of the reactive groups is an electrophilic moiety, and the second reactive group is a nucleophilic moiety. Either the nucleophilic moiety or the electrophilic moiety can be attached to the side-chain of the unnatural amino acid; the corresponding group is then attached to the saccharide moiety.

Suitable electrophilic moieties that react with nucleophilic moieties to form a covalent bond are known to those of skill in the art. In certain embodiments, such electrophilic moieties include, but are not limited to, *e.g.*, carbonyl group, a sulfonyl group, an aldehyde group, a ketone group, a

hindered ester group, a thioester group, a stable imine group, an epoxide group, an aziridine group, etc.

Suitable nucleophilic moieties that can react with electrophilic moiety are known to those of skill in the art. In certain embodiments, such nucleophiles include, for example, aliphatic or aromatic amines, such as ethylenediamine. In certain embodiments, the nucleophilic moieties include, but are not limited to, *e.g.*, -NR1-NH_2 (hydrazide), -NR1(C=O)NR2NH_2 (semicarbazide), -NR1(C=S)NR2NH_2 (thiosemicarbazide), -(C=O)NR1NH_2 (carbonylhydrazide), -(C=S) NR1NH_2 (thiocarbonylhydrazide), -(SO_2)NR1NH_2 (sulfonylhydrazide), -NR1NR2(C=O)NR3NH_2 (carbazide), NR1NR2(C=S)NR3NH_2 (thiocarbazide), -O-NH_2 (hydroxylamine), and the like, where each R1, R2, and R3 is independently H, or alkyl having 1-6 carbons, preferably H. In certain embodiments, the reactive group is a hydrazide, hydroxylamine, semicarbazide, carbohydrazide, a sulfonylhydrazide, or the like.

The product of the reaction between the nucleophile and the electrophilic moiety typically incorporates the atoms originally present in the nucleophilic moiety. Typical linkages obtained by reacting the aldehydes or ketones with the nucleophilic moieties include reaction products such as an oxime, an amide, a hydrazone, a reduced hydrazone, a carbohydrazone, a thiocarbohydrazone, a sufonylhydrazone, a semicarbazone, a thiosemicarbazone, or similar functionality, depending on the nucleophilic moiety used and the electrophilic moiety (*e.g.*, aldehyde, ketone, and/or the like) that is reacted with the nucleophilic moiety. Linkages with carboxylic acids are typically referred to as carbohydrazides or as hydroxamic acids. Linkages with sulfonic acids are typically referred to as sulfonylhydrazides or N-sulfonylhydroxylamines. The resulting linkage can be subsequently stabilized by chemical reduction.

These methods can further involve contacting the saccharide moiety with a glycosyl transferase, a sugar donor moiety, and other reactants required for glycosyl transferase activity for a sufficient time and under

appropriate conditions to transfer a sugar from the sugar donor moiety to the saccharide moiety. In certain embodiments, the method further comprises contacting the product of the glycosyl transferase reaction with at least a second glycosyl transferase and a second sugar donor moiety. In other words, certain embodiments disclosed herein provide methods in which an amino acid-linked saccharide moiety or an unnatural amino acid that includes a saccharide moiety is further glycosylated. These glycosylation steps are preferably (though not necessarily) carried out enzymatically using, for example, a glycosyltransferase, glycosidase, or other enzyme known to those of skill in the art. In some embodiments, a plurality of enzymatic steps are carried out in a single reaction mixture that contains two or more different glycosyl transferases. For example, one can conduct a galactosylating and a sialylating step simultaneously by including both sialyl transferase and galactosyl transferase in the reaction mixture.

For enzymatic saccharide syntheses that involve glycosyl transferase reactions, the recombinant cells optionally contain at least one heterologous gene that encodes a glycosyl transferase. Many glycosyl transferases are known, as are their polynucleotide sequences. See, *e.g.*, "The WWW Guide To Cloned Glycosyl transferases," (available on the World Wide Web). Glycosyl transferase amino acid sequences and nucleotide sequences encoding glycosyl transferases from which the amino acid sequences can be deduced are also found in various publicly available databases, including GenBank, Swiss-Prot, EMBL, and others.

In certain embodiments, a glycosyl transferase includes, but is not limited to, *e.g.*, a galactosyl transferase, a fucosyl transferase, a glucosyl transferase, an N-acetylgalactosaminyl transferase, an N-acetylglucosaminyl transferase, a glucuronyl transferase, a sialyl transferase, a mannosyl transferase, a glucuronic acid transferase, a galacturonic acid transferase, an oligosaccharyl transferase, and the like. Suitable glycosyl transferases include those obtained from eukaryotes or prokaryotes.

An acceptor for the glycosyl transferases will be present on the glycoprotein to be modified by methods disclosed herein. Suitable acceptors, include, for example, galactosyl acceptors such as Galβ1,4GalNAc-; Galβ1,3GalNAc-; lacto-N-tetraose-; Galβ1,3GlcNAc-; Galβ1,4GlcNAc-; Galβ1,3Ara-; Galβ1,6GlcNAc-; and Galβ1,4Glc-(lactose). Other acceptors known to those of skill in the art (see, *e.g.*, Paulson *et al.*, *J. Biol. Chem.* 253: 5617-5624, 1978). Typically, the acceptors form part of a saccharide moiety chain that is attached to the glycoprotein.

In one embodiment, the saccharide moiety comprises a terminal GlcNAc, the sugar donor moiety is UDP-GlcNAc and the glycosyl transferase is a β1-4N-acetylglucosaminyl transferase. In another embodiment, the saccharide moiety comprises a terminal GlcNAc, the sugar donor moiety is UDP-Gal and the glycosyl transferase is a β1-4-galactosyl transferase. Additional sugars can be added as well.

The glycosylation reactions include, in addition to the appropriate glycosyl transferase and acceptor, an activated nucleotide sugar that acts as a sugar donor for the glycosyl transferase. The reactions can also include other ingredients that facilitate glycosyl transferase activity. These ingredients can include a divalent cation (*e.g.*, Mg^{2+} or Mn^{2+}), materials necessary for ATP regeneration, phosphate ions, and organic solvents. The concentrations or amounts of the various reactants used in the processes depend upon numerous factors including reaction conditions such as temperature and pH value, and the choice and amount of acceptor saccharides to be glycosylated. The reaction medium may also comprise solubilizing detergents (*e.g.*, Triton or SDS) and organic solvents such as methanol or ethanol, if necessary.

Also provided by certain embodiments for modifying a glycoprotein are compositions that include a translation system which may or may not include a host cell, an external mutant tRNA, an external mutant RS, or any or all of these.

As used herein, the term "saccharide moiety" refers to natural and unnatural sugar moieties (*i.e.*, a unnaturally occurring sugar moiety, *e.g.*, a sugar moiety that is modified, *e.g.*, at one or more hydroxyl or amino positions, *e.g.*, dehydroxylated, deaminated, esterified, etc., *e.g.*, 2-deoxyGal is an example of an unnatural sugar moiety).

The term "carbohydrate" has the general formula $(CH_2O)_n$, and includes, but is not limited to, *e.g.*, monosaccharides, disaccharides, oligosaccharides and polysaccharides. Oligosaccharides are chains composed of saccharide units, which are alternatively known as sugars. Saccharide units can be arranged in any order and the linkage between two saccharide units can occur in any of approximately ten different ways. The following abbreviations are used herein: Ara=arabinosyl; Fru=fructosyl; Fuc=fucosyl; Gal=galactosyl; GalNAc=N-acetylgalactosaminyl; Glc=glucosyl; GlcNAc=N-acetylglucosaminyl; Man=mannosyl; and NeuAc=sialyl (typically N-acetylneuraminyl).

Oligosaccharides are considered to have a reducing end and a non-reducing end, whether or not the saccharide at the reducing end is in fact a reducing sugar. In accordance with accepted nomenclature, oligosaccharides are depicted herein with the non-reducing end on the left and the reducing end on the right. All oligosaccharides described herein are described with the name or abbreviation for the non-reducing saccharide (*e.g.*, Gal), followed by the configuration of the glycosidic bond (α or β), the ring bond, the ring position of the reducing saccharide involved in the bond, and then the name or abbreviation of the reducing saccharide (*e.g.*, GlcNAc). The linkage between two sugars may be expressed, for example, as 2,3; 2→3; 2-3; or (2,3). Natural and unnatural linkages (*e.g.*, 1-2; 1-3; 1-4; 1-6; 2-3; 2-4; 2-6; etc.) between two sugars are included in certain embodiments. Each saccharide is a pyranose.

The term "sialic acid" (abbreviated "Sia") refers to any member of a family of nine-carbon carboxylated sugars. The most common member of the sialic acid family is N-acetyl-neuraminic acid (2-keto-5-acetamindo-3,5-dideoxy-D-glycero-D-galactononulopyranos-1-onic acid) (often abbreviated as Neu5Ac,

NeuAc, or NANA). A second member of the family is N-glycolyl-neuraminic acid (Neu5Gc or NeuGc), in which the N-acetyl group of NeuAc is hydroxylated. A third sialic acid family member is 2-keto-3-deoxy-nonulosonic acid (KDN) (Nadano *et al.*, *J. Biol. Chem.* 261: 11550-11557, 1986; Kanamori *et al.*, *J. Biol. Chem.* 265: 21811-21819, 1990). Also included are 9-substituted sialic acids such as a 9-O-C1-C6 acyl-Neu5Ac like 9-O-lactyl-Neu5Ac or 9-O-acetyl-Neu5Ac, 9-deoxy-9-fluoro-Neu5Ac and 9-azido-9-deoxy-Neu5Ac. For review of the sialic acid family, see, *e.g.*, Varki, *Glycobiology* 2: 25-40, 1992; Sialic Acids: Chemistry, Metabolism and Function, R. Schauer, Ed. (Springer-Verlag, New York (1992). The synthesis and use of sialic acid compounds in a sialylation procedure is described in, for example, international application WO 92/16640 (entire contents incorporated herein by reference).

Donor substrates for glycosyl transferases are activated nucleotide sugars. Such activated sugars generally consist of uridine and guanosine diphosphate, and cytidine monophosphate, derivatives of the sugars in which the nucleoside diphosphate or monophosphate serves as a leaving group. Bacterial, plant, and fungal systems can sometimes use other activated nucleotide sugars.

The incorporation of an unnatural amino acid, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety, can be done to, *e.g.*, tailor changes in protein structure and/or function, *e.g.*, to change size, acidity, nucleophilicity, hydrogen bonding, hydrophobicity, accessibility of protease target sites, target access to a protein moiety, etc. Proteins that include an unnatural amino acid, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety, can have enhanced, or even entirely new, catalytic or physical properties.

For example, the following properties are optionally modified by inclusion of an unnatural amino acid, *e.g.*, an unnatural amino acid comprising

a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety into a protein: toxicity, biodistribution, structural properties, spectroscopic properties, chemical and/or photochemical properties, catalytic ability, half-life (*e.g.*, serum half-life), ability to react with other molecules, *e.g.*, covalently or noncovalently, and the like. The compositions including proteins that include at least one unnatural amino acid, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety are useful for, *e.g.*, novel therapeutics, diagnostics, catalytic enzymes, industrial enzymes, binding proteins (*e.g.*, antibodies), and *e.g.*, the study of protein structure and function. See, *e.g.*, Dougherty, *Curr. Opin. in Chem. Biol.*, 4:645-652 (2000).

In one aspect, a composition includes at least one protein with at least one, *e.g.*, at least about two, three, four, five, six, seven, eight, nine, or at least about ten or more unnatural amino acids, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety, and/or which include another unnatural amino acid. The unnatural amino acids can be the same or different, *e.g.*, there can be 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 or more different sites in the protein that comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 or more different unnatural amino acids. In another aspect, a composition includes a protein with at least one, but fewer than all, of a particular amino acid present in the protein substituted with the unnatural amino acid, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety can be attached, or an unnatural amino acid that includes a saccharide moiety. For a given protein with more than one unnatural amino acids, the unnatural amino acids can be identical or different (*e.g.*, the protein can include two or more different types of unnatural amino acids, or can include two of the same unnatural amino acid). For a given protein with more than two unnatural amino acids, the unnatural

amino acids can be the same, different, or a combination of multiple unnatural amino acids of the same kind with at least one different unnatural amino acid.

Essentially any protein (or portion thereof) that includes an unnatural amino acid, *e.g.*, an unnatural amino acid comprising a moiety where a saccharide moiety is attached, such as an aldehyde- or keto-derivatized amino acid, or an unnatural amino acid that includes a saccharide moiety (and any corresponding coding nucleic acid, *e.g.*, which includes one or more selector codons) can be produced using the compositions and methods herein. No attempt is made to identify the hundreds of thousands of known proteins, any of which can be modified to include one or more unnatural amino acid, *e.g.*, by tailoring any available mutation methods to include one or more appropriate degenerate codons in a relevant translation system. Common sequence repositories for known proteins include GenBank EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

Typically, the proteins are, *e.g.*, at least about 60%, 70%, 75%, 80%, 90%, 95%, or at least about 99% or more identical to any available protein (*e.g.*, a therapeutic protein, a diagnostic protein, an industrial enzyme, or portion thereof, and the like), and they comprise one or more unnatural amino acid.

In addition to modifying one or more amino acid residues of the protein, the protein's carbohydrate composition may be modified, *i.e.*, through glycosylation. The post-translational modification of proteins by glycosylation can affect protein folding and stability, modify the intrinsic activity of proteins, and modulate their interactions with other biomolecules. See, *e.g.*, Varki, *Glycobiology* 3: 97-130, 1993, hereby incorporated by reference in its entirety. Natural glycoproteins are often present as a population of many different glycoforms, which makes analysis of glycan structure and the study of glycosylation effects on protein structure and function difficult. Therefore, methods for the synthesis of natural and unnatural homogeneously

glycosylated proteins are needed for the systematic understanding of glycan function, and for the development of improved glycoprotein therapeutics.

One class of proteins that can be made using certain compositions and methods disclosed herein includes transcriptional modulators, enzymes, or a portion thereof. Example transcriptional modulators include genes and transcriptional modulator proteins that modulate cell growth, differentiation, regulation, or the like. Transcriptional modulators are found in prokaryotes, viruses, and eukaryotes, including fungi, plants, yeasts, insects, and animals, including mammals, providing a wide range of therapeutic targets. It will be appreciated that expression and transcriptional activators regulate transcription by many mechanisms, *e.g.*, by binding to receptors, stimulating a signal transduction cascade, regulating expression of transcription factors, binding to promoters and enhancers, binding to proteins that bind to promoters and enhancers, unwinding DNA, splicing pre-mRNA, polyadenylating RNA, and degrading RNA. Some examples of enzymes include, but are not limited to, *e.g.*, amidases, amino acid racemases, acylases, dehalogenases, dioxygenases, diarylpropane peroxidases, epimerases, epoxide hydrolases, esterases, isomerases, kinases, glucose isomerases, glycosidases, glycosyl transferases, haloperoxidases, monooxygenases (*e.g.*, p450s), lipases, lignin peroxidases, nitrile hydratases, nitrilases, proteases, phosphatases, subtilisins, transaminase, and nucleases.

Some of the polypeptides that can be modified according to certain embodiments disclosed herein are commercially available (see, *e.g.*, the Sigma BioSciences catalogue and price list), and the corresponding protein sequences and genes and, typically, many variants thereof, are well-known (see, *e.g.*, Genbank).

Examples of therapeutically relevant properties that may be manipulated or modified by any of the embodiments disclosed herein (including glycosylation and/or pegylation, and/or incorporation of non-natural amino acids) include serum half-life, shelf half-life, stability, immunogenicity,

therapeutic activity, detectability (*e.g.*, by the inclusion of reporter groups (*e.g.*, labels or label binding sites) in the unnatural amino acids, specificity, reduction of LD50 or other side effects, ability to enter the body through the gastric tract (*e.g.*, oral availability), or the like. Examples of relevant diagnostic properties include shelf half-life, stability, diagnostic activity, detectability, specificity, or the like. Examples of relevant enzymatic properties include shelf half-life, stability, specificity, enzymatic activity, production capability, or the like.

A variety of other proteins can also be modified to include one or more unnatural amino acids according to certain embodiments disclosed herein. For example, the proteins from infectious fungi, *e.g.*, *Aspergillus*, *Candida* species; bacteria, particularly *E. coli*, which serves a model for pathogenic bacteria, as well as medically important bacteria such as *Staphylococci* (*e.g.*, *aureus*), or *Streptococci* (*e.g.*, *pneumoniae*); protozoa such as sporozoa (*e.g.*, *Plasmodia*), rhizopods (*e.g.*, *Entamoeba*) and flagellates (*Trypanosoma*, *Leishmania*, *Trichomonas*, *Giardia*, etc.); viruses such as (+) RNA viruses (examples include Poxviruses *e.g.*, vaccinia; Picornaviruses, *e.g.*, polio; Togaviruses, *e.g.*, rubella; Flaviviruses, *e.g.*, HCV; and Coronaviruses), (−) RNA viruses (*e.g.*, Rhabdoviruses, *e.g.*, VSV; Paramyxovimses, *e.g.*, RSV; Orthomyxovimses, *e.g.*, influenza; Bunyaviruses; and Arenaviruses), dsDNA viruses (Reoviruses, for example), RNA to DNA viruses, *i.e.*, Retroviruses, *e.g.*, HIV and HTLV, and certain DNA to RNA viruses such as Hepatitis B.

Agriculturally related proteins such as insect resistance proteins (*e.g.*, the Cry proteins), starch and lipid production enzymes, plant and insect toxins, toxin-resistance proteins, Mycotoxin detoxification proteins, plant growth enzymes (*e.g.*, Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase, "RUBISCO"), lipoxygenase (LOX), and Phosphoenolpyruvate (PEP) carboxylase are also suitable targets for modification by certain embodiments disclosed herein.

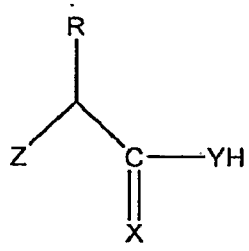
In certain embodiments, the protein or polypeptide of interest (or portion thereof) in the methods and/or compositions disclosed herein is

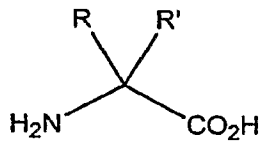
encoded by a nucleic acid. Typically, the nucleic acid comprises at least one degenerate codon, at least about two, three, four, five, six, seven, eight, nine, or at least about ten or more degenerate codons.

Thus the above-described artificial (*e.g.*, man-made, and not naturally occurring) polypeptides and polynucleotides are also features of certain embodiments disclosed herein. An artificial polynucleotide may include, *e.g.*, (a) a polynucleotide comprising a nucleotide sequence encoding an artificial polypeptide; (b) a polynucleotide that is complementary to or that encodes a polynucleotide sequence of (a); (c) a nucleic acid that hybridizes to a polynucleotide of (a) or (b) under stringent conditions over substantially the entire length of the nucleic acid; (d) a polynucleotide that is at least about 95%, preferably at least about 98% identical to a polynucleotide of (a), (b), or (c); and, (e) a polynucleotide comprising a conservative variation of (a), (b), (c), or (d).

Unnatural amino acids are generally described above. Of particular interest for making glycoproteins as described herein are unnatural amino acids in which R in Formula I includes a moiety that can react with a reactive group that is attached to a saccharide moiety, to link the saccharide moiety to a protein that includes the unnatural amino acid. Suitable R groups include, for example, keto-, azido-, hydroxyl-, hydrazine, cyano-, halo-, aminooxy-, alkenyl, alkynyl, carbonyl, ether, thiol, seleno-, sulfonyl-, borate, boronate, phospho, phosphono, phosphine, heterocyclic, enone, imine, aldehyde, ester, thioacid, thioester, hindered ester, hydroxylamine, amine, and the like, or any combination thereof. In some embodiments, the unnatural amino acids have a photoactivatable cross-linker.

In addition to unnatural amino acids that contain novel side chains, unnatural amino acids also optionally comprise modified backbone structures, *e.g.*, as illustrated by the structures of Formula II and III:





Formula II Formula III

wherein Z typically comprises OH, NH_2, SH, NH-R', or S-R'; X and Y, which can be the same or different, typically comprise S or O, and R and R', which are optionally the same or different, are typically selected from the same list of constituents for the R group described above for the unnatural amino acids having Formula I as well as hydrogen. For example, unnatural amino acids disclosed herein are optionally comprise substitutions in the amino or carboxyl group as illustrated by Formulas II and III. Unnatural amino acids of this type include, but are not limited to, α-hydroxy acids, α-thioacids α-aminothiocarboxylates, *e.g.*, with side chains corresponding to the common twenty natural amino acids or unnatural side chains. In addition, substitutions at the α-carbon optionally include L, D, or α-α-disubstituted amino acids such as D-glutamate, D-alanine, D-methyl-O-tyrosine, aminobutyric acid, and the like. Other structural alternatives include cyclic amino acids, such as proline analogues as well as 3-, 4-, 6-, 7-, 8-, and 9-membered ring proline analogues, β and γ amino acids such as substituted β-alanine and γ-amino butyric acid.

For example, many unnatural amino acids are based on natural amino acids, such as tyrosine, glutamine, phenylalanine, and the like. Tyrosine analogs include para-substituted tyrosines, ortho-substituted tyrosines, and meta substituted tyrosines, wherein the substituted tyrosine comprises an acetyl group, a benzoyl group, an amino group, a hydrazine, an hydroxyamine, a thiol group, a carboxy group, an isopropyl group, a methyl group, a C6-C20 straight chain or branched hydrocarbon, a saturated or unsaturated hydrocarbon, an O-methyl group, a polyether group, a nitro group, or the like. In addition, multiply substituted aryl rings are also contemplated. Glutamine

analogs include, but are not limited to, α-hydroxy derivatives, γ-substituted derivatives, cyclic derivatives, and amide substituted glutamine derivatives. Example phenylalanine analogs include, but are not limited to, meta-substituted, ortho-substituted, and/or para-substituted phenylalanines, wherein the substituent comprises a hydroxy group, a methoxy group, a methyl group, an allyl group, an aldehyde or keto group, or the like.

Specific examples of unnatural amino acids include, but are not limited to, p-acetyl-L-phenylalanine, O-methyl-L-tyrosine, an L-3-(2-naphthyl)alanine, a 3-methyl-phenylalanine, an O-4-allyl-L-tyrosine, a 4-propyl-L-tyrosine, a tri-O-acetyl-GlcNAcβ-serine, β-O-GlcNAc-L-serine, a tri-O-acetyl-GalNAc-α-threonine, an α-GalNAc-L-threonine, an L-Dopa, a fluorinated phenylalanine, an isopropyl-L-phenylalanine, a p-azido-L-phenylalanine, a p-acyl-L-phenylalanine, a p-benzoyl-L-phenylalanine, an L-phosphoserine, a phosphonoserine, a phosphonotyrosine, a p-iodo-phenylalanine, a p-bromophenylalanine, a p-amino-L-phenylalanine, an isopropyl-L-phenylalanine, those listed below, or elsewhere herein, and the like.

Unnatural amino acids suitable for use in some methods disclosed herein also include those that have a saccharide moiety attached to the amino acid side chain. In one embodiment, an unnatural amino acid with a saccharide moiety includes a serine or threonine amino acid with a Man, GalNAc, Glc, Fuc, or Gal moiety. Examples of unnatural amino acids that include a saccharide moiety include, but are not limited to, *e.g.*, a tri-O-acetyl-GlcNAcβ-serine, a β-O-GlcNAc-L-serine, a tri-O-acetyl-GalNAc-α-threonine, an α-GalNAc-L-threonine, an O-Man-L-serine, a tetra-acetyl-O-Man-L-serine, an O-GalNAc-L-serine, a tri-acetyl-O-GalNAc-L-serine, a Glc-L-serine, a tetraacetyl-Glc-L-serine, a fuc-L-serine, a tri-acetyl-fuc-L-serine, an O-Gal-L-serine, a tetra-acetyl-O-Gal-L-serine, a beta-O-GlcNAc-L-threonine, a tri-acetyl-beta-GlcNAc-L-threonine, an O-Man-L-threonine, a tetra-acetyl-O-Man-L-threonine, an O-GalNAc-L-threonine, a tri-acetyl-O-GalNAc-L-threonine, a Glc-L-threonine, a tetraacetyl-Glc-L-threonine, a fuc-L-threonine, a tri-acetyl-fuc-L-

threonine, an O-Gal-L-threonine, a tetra-acetyl-O-Gal-L-serine, and the like. Certain embodiments also include unprotected and acetylated forms of the above.

In some embodiments, the design of unnatural amino acids is biased by known information about the active sites of synthetases, *e.g.*, external mutant tRNA synthetases used to aminoacylate an external mutant tRNA. For example, three classes of glutamine analogs are provided, including derivatives substituted at the nitrogen of amide (1), a methyl group at the γ-position (2), and a N-Cγ-cyclic derivative (3). Based upon the x-ray crystal structure of *E. coli* GlnRS, in which the key binding site residues are homologous to yeast GlnRS, the analogs were designed to complement an array of side chain mutations of residues within a 10 Å shell of the side chain of glutamine, *e.g.*, a mutation of the active site Phe233 to a small hydrophobic amino acid might be complemented by increased steric bulk at the Cγ position of Gln.

For example, N-phthaloyl-L-glutamic 1,5-anhydride (compound number 4 in FIG. 23 of WO 02/085923) is optionally used to synthesize glutamine analogs with substituents at the nitrogen of the amide. See, *e.g.*, King and Kidd, *J. Chem. Soc.*, 3315-3319, 1949; Friedman and Chatterrji, *J. Am. Chem. Soc.* 81, 3750-3752, 1959; Craig *et al.*, *J. Org. Chem.* 53, 1167-1170, 1988; and Azoulay *et al.*, *Eur. J. Med. Chem.* 26, 201-5, 1991. The anhydride is typically prepared from glutamic acid by first protection of the amine as the phthalimide followed by refluxing in acetic acid. The anhydride is then opened with a number of amines, resulting in a range of substituents at the amide. Deprotection of the phthaloyl group with hydrazine affords a free amino acid as shown in FIG. 23 of WO 2002/085923.

Substitution at the γ-position is typically accomplished via alkylation of glutamic acid. See, *e.g.*, Koskinen and Rapoport, *J. Org. Chem.* 54, 1859-1866, 1989. A protected amino acid, *e.g.*, as illustrated by compound number 5 in FIG. 24 of WO 02/085923, is optionally prepared by first alkylation

of the amino moiety with 9-bromo-9-phenylfluorene (PhflBr) (see, *e.g.*, Christie and Rapoport, *J. Org. Chem.* 1989, 1859-1866, 1985) and then esterification of the acid moiety using O-tert-butyl-N,N'-diisopropylisourea. Addition of $KN(Si(CH_3)_3)_2$ regioselectively deprotonates at the α-position of the methyl ester to form the enolate, which is then optionally alkylated with a range of alkyl iodides. Hydrolysis of the t-butyl ester and Phfl group gave the desired γ-methyl glutamine analog (Compound number 2 in FIG. 24 of WO 02/085923).

Certain other embodiments include an immunoconjugate comprising an antibody (or its functional fragment) specific for a target (*e.g.*, a target cell), the antibody (or fragment or functional equivalent thereof) conjugated, at specific, pre-determined positions, with two or more therapeutic molecules, wherein each of the positions comprise an unnatural amino acid. In certain embodiments, the antibody fragments are $F(ab')_2$, Fab', Fab, or Fv fragments.

In certain embodiments, the two or more therapeutic molecules are the same. In certain embodiments, the two or more therapeutic molecules are different. In certain embodiments, the therapeutic molecules are conjugated to the same unnatural amino acids. In certain embodiments, the therapeutic molecules are conjugated to different unnatural amino acids.

In certain embodiments, the nature or chemistry of the unnatural amino acid / therapeutic molecule linkage allows cleavage of the linkage under certain conditions, such as mild or weak acidic conditions (*e.g.*, about pH 4-6, preferably about pH5), reductive environment (*e.g.*, the presence of a reducing agent), or divalent cations, and is optionally accelerated by heat.

In certain embodiments, the therapeutic molecule is conjugated to an antibody through a linker / spacer (*e.g.*, one or more repeats of methylene (-CH_2-), methyleneoxy (-CH_2-O-), methylenecarbonyl (-CH_2-CO-), amino acids, or combinations thereof).

Multiprotein complexes

Unnatural amino acids can also be used to join two or more proteins or protein sub-units with unique functionalities. For example, bispecific antibodies may be generated by linking two antibodies (or functional parts thereof or derivatives thereof, such as Fab, Fab', Fd, Fv, scFv fragments, etc.) through unnatural amino acids incorporated therein.

Thus certain embodiments herein provide methods for synthesis of multi-protein conjugates. These methods involve, in some embodiments, incorporating into a first protein (*e.g.*, a first antibody) a first unnatural amino acid that comprises a first reactive group; and contacting the first protein with a second protein (*e.g.*, a second antibody) comprising a second unnatural amino acid that comprises a second reactive group, wherein the first reactive group reacts with the second reactive group, thereby forming a covalent bond that attaches the second protein to the first protein.

The first reactive group is, in some embodiments, an electrophilic moiety (*e.g.*, a keto moiety, an aldehyde moiety, and/or the like), and the second reactive group is a nucleophilic moiety. In some embodiments, the first reactive group is a nucleophilic moiety and the second reactive group is an electrophilic moiety (*e.g.*, a keto moiety, an aldehyde moiety, and/or the like). For example, an electrophilic moiety is attached to the unnatural amino acid of the first Ab, and the nucleophilic moiety is attached to the unnatural amino acid of the second Ab.

Different functional domains of different proteins may be linked together through similar fashion to create novel proteins with novel functions (*e.g.*, novel transcription factors with unique combination of DNA binding and transcription activation domains; novel enzymes with novel regulatory domains, etc.).

pH-Sensitive Binding

Many protein interactions are pH-sensitive, in the sense that binding affinity of one protein for its usual binding partner may change as environmental pH changes. For example, many ligands (such as insulin, interferons, growth hormone, etc.) bind their respective cell-surface receptors to elicit signal transduction. The ligand-receptor complex will then be internalized by receptor-mediated endocytosis, and go through a successive series of more and more acidic endosomes. Eventually, the ligand-receptor interaction is weakened at a certain acidic pH (*e.g.*, about pH 5.0), and the ligand dissociates from the receptor. Some receptors (and perhaps some ligands) may be recycled back to cell surface. There, they may be able to bind their respective normal binding partners.

If the pH-sensitive binding can be modulated such that the ligand-receptor complex can be dissociated at a relatively higher pH, then certain ligands may be dissociated earlier from their receptors, and become preferentially recycled to cell surface rather than be degraded. This will result in an increased in vivo half-life of such ligands, which might be desirable since less insulin may be needed for the same (or better) efficacy in diabete patients. In other situations, it might be desirable to modulate the pH-sensitive binding by favoring binding at a lower pH.

For example, monoclonal antibodies are generally very specific for their targets. However, in many applications, such as in cancer therapy, they tend to elicit certain side effects by, for example, binding to non-tumor tissues. One reason could be that the tumor targets against which monoclonal antibodies are raised are not specifically expressed on tumor cells, but are also expressed (although may be in smaller numbers) on some healthy cells. Such side effects are generally undesirable, and there is a need for antibodies with an improved specificity.

The pH of human blood is highly regulated and maintained in the range of about 7.6-7.8. On the other hand, tumor cells have an extracellular pH

of 6.3-6.5, due to the accumulation of metabolic acids that are inefficiently cleared because of poor tumor vascularization. If the interaction between a tumor antigen and its therapeutic antibody can be modulated such that at low pH, the binding is favored, the tumor-antibody may have an added specificity / affinity / selectivity for those tumor antigens, even though the same tumor antigens are also occasionally found on normal tissues.

In fact, such modified antibodies may be desirable not only for cancer therapy, but also desirable for any antigen-antibody binding that may occur at a lower-than-normal level of pH.

General Techniques

The practice of the embodiments disclosed herein will employ, unless otherwise indicated, conventional techniques of molecular biology, cell biology, cell culture, microbiology and recombinant DNA, which are within the skill of the art. Such techniques are explained fully in the literature. See, for example, *Molecular Cloning: A Laboratory Manual*, 2nd Ed., ed. By Sambrook, Fritsch and Maniatis (Cold Spring Harbor Laboratory Press: 1989); *DNA Cloning*, Volumes I and II (D. N. Glover ed., 1985); *Oligonucleotide Synthesis* (M. J. Gait ed., 1984); Mullis *et al.*; U.S. Patent No: 4,683,195; *Nucleic Acid Hybridization* (B. D. Hames & S. J. Higgins eds. 1984); *Transcription And Translation* (B. D. Hames & S. J. Higgins eds. 1984); B. Perbal, *A Practical Guide To Molecular Cloning* (1984); the treatise, *Methods In Enzymology* (Academic Press, Inc., N.Y.); *Methods In Enzymology*, Vols. 154 and 155 (Wu *et al.* eds.), *Immunochemical Methods In Cell And Molecular Biology* (Mayer and Walker, eds., Academic Press, London, 1987).

Furthermore, general texts disclosing general cloning, mutation, cell culture and the like, include Berger and Kimmel, *Guide to Molecular Cloning Techniques, Methods in Enzymology* vol. 152 Academic Press, Inc., San Diego, Calif. (Berger); Sambrook *et al.*, *Molecular Cloning—A Laboratory Manual* (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor,

N.Y., 2000 ("Sambrook") and *Current Protocols in Molecular Biology*, F. M. Ausubel *et al.*, eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2002) ("Ausubel")) are all hereby incorporated by reference in their entireties. These texts describe mutagenesis, the use of vectors, promoters and many other relevant topics related to, *e.g.*, the generation of external mutant tRNA, external mutant synthetases, and pairs thereof.

Various types of mutagenesis are used in certain embodiments, *e.g.*, to produce novel sythetases or tRNAs. They include but are not limited to site-directed (such as through use of Amber, Ochre, Umber or other stop codon), via wobble codon mutagenesis, random point mutagenesis, homologous recombination (DNA shuffling), mutagenesis using uracil containing templates, oligonucleotide-directed mutagenesis, phosphorothioate-modified DNA mutagenesis, mutagenesis using gapped duplex DNA or the like. Additional suitable methods include point mismatch repair, mutagenesis using repair-deficient host strains, restriction-selection and restriction-purification, deletion mutagenesis, mutagenesis by total gene synthesis, double-strand break repair, and the like. Mutagenesis, *e.g.*, involving chimeric constructs, are also included in certain embodiments. In one embodiment, mutagenesis can be guided by known information of the naturally occurring molecule or altered or mutated naturally occurring molecule, *e.g.*, sequence, sequence comparisons, physical properties, crystal structure or the like.

The above texts and examples found herein describe these procedures as well as the following publications and references cited within: Sieber, *et al.*, *Nature Biotechnology*, 19:456-460 (2001); Ling *et al.*, *Approaches to DNA mutagenesis: an overview*, *Anal. Biochem.* 254(2): 157-178 (1997); Dale *et al.*, *Methods Mol. Biol.* 57:369-374 (1996); I. A. Lorimer, I. Pastan, *Nucleic Acids Res.* 23, 3067-8 (1995); W. P. C. Stemmer, *Nature* 370, 389-91 (1994); Arnold, *Curr. Opin. in Biotech.* 4:450-455 (1993); Bass *et al.*, *Science* 242:240-245 (1988); Fritz *et al.*, *Nucl. Acids Res.* 16: 6987-6999 (1988);

Kramer *et al.*, *Nucl. Acids Res.* 16: 7207 (1988); Sakamar and Khorana, *Nucl. Acids Res.* 14: 6361-6372 (1988); Sayers *et al.*, *Nucl. Acids Res.* 16:791-802 (1988); Sayers *et al.*, *Nucl. Acids Res.* 16: 803-814 (1988); Carter, *Methods in Enzymol.* 154: 382-403 (1987); Kramer & Fritz *Methods in Enzymol.* 154:350-367 (1987); Kunkel, *Nucleic Acids & Mol. Biol.* (Eckstein, F. and Lilley, D. M. J. eds., Springer Verlag, Berlin)) (1987); Kunkel *et al.*, *Methods in Enzymol.* 154, 367-382 (1987); Zoller & Smith, *Methods in Enzymol.* 154:329-350 (1987); Carter, *Biochem. J.* 237:1-7 (1986); Eghtedarzadeh & Henikoff, *Nucl. Acids Res.* 14: 5115 (1986); Mandecki, *PNAS,* USA, 83:7177-7181 (1986); Nakamaye & Eckstein, *Nucl. Acids Res.* 14: 9679-9698 (1986); Wells *et al.*, *Phil. Trans. R. Soc. Lond.* A 317: 415-423 (1986); Botstein & Shortle, *Science* 229:1193-1201(1985); Carter *et al.*, *Nucl. Acids Res.* 13: 4431-4443 (1985); Grundström *et al.*, *Nucl. Acids Res.* 13: 3305-3316 (1985); Kunkel, *PNAS,* USA 82:488-492 (1985); Smith, *Ann. Rev. Genet.* 19:423-462(1985); Taylor *et al.*, *Nucl. Acids Res.* 13: 8749-8764 (1985); Taylor *et al.*, *Nucl. Acids Res.* 13: 8765-8787 (1985); Wells *et al.*, *Gene* 34:315-323 (1985); Kramer *et al.*, *Nucl. Acids Res.* 12: 9441-9456 (1984); Kramer *et al.*, *Cell* 38:879-887 (1984); Nambiar *et al.*, *Science* 223: 1299-1301 (1984); Zoller & Smith, *Methods in Enzymol.* 100:468-500 (1983); and Zoller & Smith, *Nucl Acids Res.* 10:6487-6500 (1982). Additional details on many of the above methods can be found in Methods in Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Oligonucleotides, *e.g.*, for use in mutagenesis in certain embodiments, *e.g.*, mutating libraries of synthetases, or altering tRNAs, are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers, *Tetrahedron Letts.* 22(20):1859-1862, (1981) *e.g.*, using an automated synthesizer, as described in Needham-VanDevanter *et al.*, *Nucl Acids Res.*, 12:6159-6168 (1984).

In addition, essentially any nucleic acid can be custom or standard ordered from any of a variety of commercial sources, such as The

Midland Certified Reagent Company, The Great American Gene Company, ExpressGen Inc., Operon Technologies Inc. (Alameda, Calif.) and many others.

All embodiments described herein are intended to be able to be combined with one or more other embodiments, even for those described under different sections of the disclosure.

All of the above U.S. patents, U.S. patent application publications, U.S. patent applications, foreign patents, foreign patent applications and non-patent publications referred to in this specification and/or listed in the Application Data Sheet, are incorporated herein by reference, in their entirety.

EXAMPLES

These examples illustrate the incorporation of an amino acid analog in proteins at positions encoded by codons which normally specifically encode phenylalanine (Phe) or specifically encode tryptophan (Trp). A schematic diagram is shown in Figure 1. Similar approaches can be used for any other analogs.

Phe is encoded by two codons, UUC and UUU. Both codons are read by a single tRNA, which is equipped with the anticodon sequence GAA. The UUC codon is therefore recognized through standard Watson-Crick base-pairing between codon and anticodon; UUU is read through a G-U wobble base-pair at the first position of the anticodon (Crick, *J. Mol. Biol.* 19: 548, 1966; Soll and RajBhandary, *J. Mol. Biol.* 29: 113, 1967). Thermal denaturation of RNA duplexes has yielded estimates of the Gibbs free energies of melting of G-U, G-C, A-U, and A-C basepairs as 4.1, 6.5, 6.3, and 2.6 kcal/mol, respectively, at 37°C. Thus the wobble basepair, G-U, is less stable than the Watson-Crick basepair, A-U. A modified $tRNA^{Phe}$ outfitted with the AAA anticodon ($tRNA^{Phe}_{AAA}$) was engineered to read the UUU codon, and was predicted to read such codons faster than wild-type $tRNA^{Phe}_{GAA}$.

Murine dihydrofolate reductase (mDHFR), which contains nine Phe residues, was chosen as the test protein. The expression plasmid pQE16

encodes mDHFR under control of a bacteriophage T5 promoter; the protein is outfitted with a C-terminal hexahistidine (HIS_6) tag to facilitate purification via immobilized metal affinity chromatography.

The modified yeast PheRS (mu-yPheRS) was prepared by introduction of a Thr415Gly or Thr415Ala mutation in the α-subunit of the synthetase (Datta el al., *J. Am. Chem. Soc.* 124: 5652, 2002). The kinetics of activation of Nal and Phe by mu-yPheRS were analyzed *in vitro* via the adenosine triphosphate-pyrophosphate exchange assay. The specificity constant (k_{cat} / K_M) for activation of Nal by mu-yPheRS was found to be 1.55 × 10^{-3} (s^{-1} M^{-1}), 8-fold larger than that for Phe. Therefore, when the ratio of Nal to Phe in the culture medium is high, ytRNA$^{Phe}_{AAA}$ should be charged predominantly with Nal. In addition, the T415G mutant was generated by four-primer mutagenesis.

Both *E.coli* and yeast synthetases are $\alpha_2\beta_2$ hetero-tetramers and the molecular weights for each subunit are rather different α(ePheRS) = 37 kDA;α (yPheRS)= 57 kDa; β(ePheRS) = 87 kDa; and β(yPheRS) = 67.5, all approximately.

Thus, the following examples are provided as way of illustration and not by way of limitation.

EXAMPLE 1

In order to alter the capability of a yeast aminoacyl tRNA synthetase, the yPheRS gene was amplified from template plasmid pUC-ASab2 encoding alpha and beta subunits of the PheRS gene. The amplification was conducted with a 14 base pair intergenic sequence containing a translational reinitiation site upstream of the ATG start code of the beta subunit gene.

The following oligo primers were used for the PCR: 5'-CGA TTT TCA CAC AGG ATC CAG ACC ATG ATT CTA G-3' (SEQ ID NO:7) (primer 1 with restriction site *BamH*I) and 5'-GAC GGC CAG TGA ATT CGA GCT CGG TAC-3' (SEQ ID NO: 8) (primer 2 with restriction site *Kpn*I). The resulting DNA

product was introduced into the *BamH*I and *Kpn*I sites of pQE32 to give pQE32-yFRS. The mutant yPheRS polynucleotide was generated by using primer mutagenesis by standard techniques.

Briefly, two complementary oligonucleotides: 5'-CTA CCT ACA ATC CTT ACG GCG AGC CAT CAA TGG AAA TC-3' (SEQ ID NO:9) (primer 3) and 5'-GAT TTC CAT TGA TGG CTC GCC GTA AGG ATT GTA GGT AG-3' (SEQ ID NO: 10) (primer 4) were synthesized to carry the specific mutation at position 415 of the alpha subunit of the yPheRS polynucleotide.

EXAMPLE 2

The plasmid pQE32-yFRS, and pQE32-T415G, pQE32-T415A were each transformed into *E.coli* host cell strain BLR (from NOVAGEN®) to form expression strains BLR(pQE32-yFRS_ and BLR(pQE32-T415G). Cells were grown in LB media, to a concentration of 0.6 at OD 600. Expression was then induced with 1 mM IPTG for 4 hours. Cells were harvested and polypeptides were purified by way of a nickel-nitrilotriacetic acid affinity column under native conditions according to the manufacturer's protocol (QIAGEN®). The imidazole in the elution buffer was removed by desalting column, and polypeptides were eluted into a buffer containing 50 mM Tris-HCl (pH = 7.5), 1 mM DTT. Aliquots of polypeptides were stored in –80°C with 50% glycerol.

EXAMPLE 3

The amino acid dependent ATP-PP_i exchange reaction was used to evaluate the activation of non-natural amino acids by yPheRS. The assay was performed in 200 mincroliters of reaction buffer containing 50 mM HEPES (pH=7.6), 20 mM $MgCl_2$, 1 mM DTT, 2 mM ATP and 2 mM [^{32}P]-PP_i with specific activity of 0.2—0.5 TBq/mol. Depending on the activity of the various non-natural amino acids with the synthetase, the amino acid concentration varied from 10 microM to 5 mM and enzyme concentration varied from 10 nM to

100 nM. Aliquots of 20 microliters were removed from the reaction solution at various time points and quenched into 500 microliters of buffer solution containing 200 mM $NaPP_i$, 7% w/v $HClO_4$ and 3% w/v activated charcoal. The charcoal was spun down and washed twice with 500 microliters of 10 mM $NaPP_i$ and 0.5% $HclO_4$ solution. The radio-labeled ATP absorbed into the charcoal was quantified via liquid scintillation methods. The specificity constants were calculated by nonlinear regression fit of the data to a Michaelis Menten model. The kinetic parameters for the ATP-PPi exchange of amino acids by the yPheRS (T415G), wild type yPheRS, and yPheRS_naph variant are shown in the table below.

Amino Acid	Enzyme	Km (μM)	Kcat (s^{-1})	Kcat/Km ($M^{-1}s^{-1}$)	Kcat/Km (relative
Phe	T415G	55+/-14	0.202+/-0.11	3512+/-1134	1[a]
Trp	T415G	2.83+/-1.6	0.153+/-0.003	63190+/-34590	18[a]
2Nal	T415G	7.03+/-0.14	0.208+/-0.04	29535+/-5848	8.4[a]
Phe	wild type	3.85+/-0.99	0.181+/-0.011	50994+/-22655	15[a]
Phe	naph	11010+/-2688	0.0095+/-0.0021	0.855+/-0.007	1[b]
Trp	naph	1424+/-597	0.0035+/-0.0009	2.52+/-0.44	2.9[b]
2Nal	naph	2030+/-691	0.030+/-0.018	14.54+/-4.22	17[b]

EXAMPLE 4

The expression plasmid, pQE16 (QIAGEN®) was used with marker polypeptide murine dihydrofolate reductase (mDHFR) with a C-terminal hexa-histidine tag gene under the control of a bacteriophage T5 promoer and t_0 terminator.

An Amber codon (TAG) was placed at the 38th position of mDHFR using a QUICK-CHANGE® mutagenesis kit. Two complementary oligo primers (5'-CCG CTC AGG AAC GAG TAG AAG TAC TTC CAA AGA ATG-3' (SEQ ID NO: 11) and 5'-CAT TCT TTG GAA GTA CTT CTA CTC GTT CCT GAG CGG-

3' (SEQ ID NO: 12)) were used to produce pQE16am. The mutant yPheRS gene T415G was amplified from pQE32-T415G and a constitutive *tac* promoter with an abolished *lac* repressor binding site was added upstream from the start codon of the gene.

The entire expression cassette of T415G was inserted into *Pvu*II site of pQE16 to yield pQE16am-T415G. The mutant yeast suppressor tRNA (*mut*RNA^{Phe}(CUA)) was constitutively expressed under control of *lpp* promoter. The expression cassette of *mut*RNA^{Phe}(CUA) was inserted into repressor plasmid pREP4 to form pREP4-tRNA using known methods.

A phenylalanine (Phe) auxotrophic bacterial strain, AF (K10, Hfr(Cavalli) *pheS13rel-1 tonA*22 thi T2^{R} *pheA*18) was used as a host strain. A Phe/Trp double auxotrophic double strain, AFW (K10, Hfr(Cavalli) *pheS13rel*-1 *tonA*22 thi T2^{R} *phe*A18, *trpB*114) and a Phe/Trp/Lys triple auxotrophic strain AFWK (K10, Hfr(Cavalli) *pheS13rel*-1 *tonA*22 thi T2^{R} *phe*A18, *trp*B114, *lysA*) were prepared by P1 phage-mediated transduction with *trpB*::*Tn10* and *lysA*::*Tn10* transposons.

EXAMPLE 5

The auxotrophic host cell strains AF, AFW, and AFWK were each transformed with plasmid pQE16am-T415G and pREP4-tRNA to yield expression strains AF[pQE16am-T415G/pREP4-tRNA] and AWF[pQE16am-T415G/pREP4-tRNA], respectively. The cells were grown in M9 minimal medium supplemented with glucose, thiamin, $MgSO_4$, $CaCl_2$, 20 amino acids (20 mg/L), antibiotics (kanamycin and ampicillin). When cells reached an OD600 reading of 1.0, they were sedimented by centrifugation, washed twice with cold 0.9% NaCl, and shifted to supplemented M9 medium containing 17 amino acids (20 mg/L), 3 mM non-natural amino acid of interest, and the indicated concentrations of Phe, Trp, and Lys. Protein expression was induced by adding IPTG (1 mM). After 4 hours, cells were pelleted and the protein was

purifed by way of a C-terminal hexa-Histidine tag and a Nickel-NTA spin column according to manufacturer's directions. (QIAGEN®).

EXAMPLE 6

Mutant mDHFR was purified under denaturing conditions and eluted with standard buffer (8 M urea, 100 mM NaH_2PO_4, 10 mM Tris, pH 4.5). The polypeptides were trypsin digested with 10 microliters of the solution diluted into 90 microliters of 75 mM $(NH_4)_2CO_3$ and the pH was adjusted to 8. Two microliters of modified trypsin (0.2 micrograms/microliter) was added and the sample was incubated at room temperature overnight. The polypeptides were endoproteinase digested with Lys-C, 10 microliters of solution diluted in 90 microliters of 25 mM Tris-HCl, pH 8 and 1 mM EDTA. Next, 2 microliters of Lys-C (0.2 micrograms/microliter; CALBIOCHEM®) was added and the reaction was incubated at 37° for 10 hours. The digestion reaction was stopped by adding 2 microliters of trifluoroacetic acid (TFA). The solution was purified by way of $ZIPTIP_{C18}$® (MILLIPORE®) and the digested peptides were eluted with 3 microliters of 50% CH_3CN, 0.1% TFA. One microliter was used for matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) analysis with alpha-cyano-4-hycroxycinnamic acid and 2, 5-dihydroxybenzoic acid as the matrix. The analysis was performed using a PERSEPTIVE BIOSYSTEMS® Voyager DE PRO MALDI-TOF mass spectrometer in linear and positive ion modes.

LC-MS/MS analysis of protease-digested peptides was conducted on FINNIGAN® LCQ ion trap mass spectrometry with HPLC pump and ESI probe. Tandem mass sequencing was carried out by fragmentation of the precursor ion with m/z corresponding to protease-digested fragment including the residue at position 38 of mutant mDHFR.

EXAMPLE 7

Plasmids were constructed for wild type yPheRS and yPheRS (T415G) as described in Example 1 herein. In addition, the *E.coli lysS* gene was amplified by PCR from template plasmid pXLLysKS1, using the following primers: 5'-GCA CTG ACC ATG GCT GAA CAA CAC GCA CAG-3' (SEQ ID NO: 13) (with *Nco*I restriction site) and 5'-GGA CTT CGG ATC CTT TCT GTG GGC GCA TCG C-3' (SEQ ID NO: 14) (with *BamH*I restriction site). The resulting DNA was introduced into the *Nco*I and *BamH*I sites of pQE60 to yield pQE60-eLysS. The cloned enzymes contain N-terminal or C-terminal hexaHistidine tags to facilitate protein purification.

At the first two reactions, (primer 1 and primer 4) and (primer 2 and primer 3) were added into individual tubes and two DNA fragments were generated from these two PCR reactions. With the mixture of two reaction products and additional outside primers, a 3400bp DNA fragment was obtained. The fragment was purified by standard methods and digested with *BamH*I and *Kpn*I and inserted into pQE32 to produce pQE32-T415G. The cloned PheRS enzymes contained an N-terminal known hexa-histidine sequence tag for purification. The entire yPheRS gene was DNA sequenced for verification.

EXAMPLE 8

The plasmid PQE32-T415A, and pQE60-eLysS were individually co-transformed with a repressor plasmid pREP4 into an *E.coli* strain BLR to form expression strains BLR (pQE32-yFRS), BLR (pQE32-T415G), BLR (pQE32-T415A) and BLR (pQE60-eLysS). Overexpression was conducted in 2x YT media with 100 micrograms/mL of ampicillin and 35 micrograms/mL of kanamycin. At OD 600 = 0.6, expression of yPheRS variants and *E.coli* lysyl-tRNA synthetase encoded by the *lysS* gene (eLysS) were induced with 1 mM IPTG. After 4 hour expression, cells were harvested and proteins were purified over a nickel-nitrilotriacetic acid affinity column under native conditions

according to the manufacturer's protocol (QIAGEN®). The imidazole in the elution buffer ewas removed by desalting column and polypeptides were eluted into a buffer containing 50 mM Tris-HCl (pH 7.5), 1 mM DTT. Aliquots of polypeptides were stored in -80^0C with 50% glycerol. Concentrations of yPheRS variants and eLysS were determined by UV absorbance at 280 nm.

EXAMPLE 9

The peptide38 (residues 26-39; NGDLPWPPLRNEamber codonK) (SEQ ID NO: 15) contains the amber codon at position number 38. Peptides (K38S, K38L), Peptide W38 and Peptide pBrF (Z)38 were separated and detected by MS. Polypeptides were synthesized in triple auxotrophic host cells with (a) $tRNA^{Phe}_{CUA}$ and yPheRS (T415G); (b) $tRNA^{Phe}_{CUA\_UG}$ and yPheRS (T415G); (c) $tRNA^{Phe}_{CUA}$ and yPheRS (T415A); (d) $tRNA^{Phe}_{CUA\_UG}$ and yPheRS (T415A) or (e) in a single auxotrophic strain with $tRNA^{Phe}_{CUA\_UG}$ and yPheRS (T415A). The expression minimal media were supplemented with 6.0 mM pBrF, 0.01 mM Trp, 1.0 mM Lys, 0.03 mM Phe (a and b) or 0.01 mM Phe (c, d and e) and 25 mg/L of 17 amino acids, results are shown in Figure 9.

EXAMPLE 10

The amino acid-dependent ATP-PP_i exchange reaction was used to evaluate the activation of amino acid analogs by yPheRS as described in the above Examples. Briefly, a 200 microliter aliquot of reaction buffer contains 50 mM HEPES (pH 7.6), 20 nM $MgCl_2$, 1 mM DTT, 2 mM ATP, and 2mM ^{32}P-pyrophosphate (PPi) with specific activity of 10-50 Ci/mol. Depending on the activyt of analogs by the synthetase, the amino acid concentration varied from 10 microM to 2.5 mM and enzyme concentration varied from 10nM to 100 nM. Aliquots of 20 microliters were removed from the reaction solution at various time points and quenched into 500 microliters of buffer solution containing 200 mN $NaPP_i$, 7% w/v $HclO_4$, and 3% w/v activated charcoal. The charcoal was

spun down and washed twice with 500 microliters of 10 mM $NaPP_i$ and 0.5% $HClO_4$ solution. The radio-labeled ATP absorbed into the charcoal was quantified via liquid scintillation methods. The specificity constants were calculated by non linear regression fit for the data to a Michaelis Menten model.

The results of the kinetic parameters are shown in Table I. Substitution at the indole ring (especially at the 6th position) was highly favorable for some analogs (8-10).

Table I: Kinetic Parameters for ATP-PPi exchange of exemplary amino acids (1-11) by the external mutant yeast PheRS.

Amino Acid	Enzyme	Km (μM)	kcat (s^{-1})	kcat/Km ($M^{-1}s^{-1}$)	kcat/Km (relative)*
1	T415G	264+/-42	0.05+/-0.002	184+/-30	1
2	T415G	22+/-3	0.03+/-0.001	1,538+/-228	8
3	T415G	12+/-2	0.05+/-0.001	4,365+/-797	24
4	T415G	11+/-3	0.05+/-0.002	4,558+/-1,186	25
5	T415G	757+/-149	0.4+/-0.003	48+/-10	1\4
6	T415G	20+/-5	0.30+/-0.006	15,000+/-4,063	82
7	T415G	27+/-2	0.04+/-0.001	1,550+/-125	8
8	T415G	20+/-8	0.20+/-0.018	10,256+/-4,562	56
9	T415G	8+/-4	0.55+/-0.097	70,876+/-34,843	385
10	T415G	31+/-18	0.06+/-0.005	1,939+/-1,149	10
11	T415G	94+/-50	0.05+/-0.006	533+/-293	3
1	T415G	68+/-20	0.52+/-0.093	7,627+/-2,664	41

Where the amino acids are indicated as in Figure 2.

EXAMPLE 11

The mutant yeast amber suppressor tRNA ($ytRNA^{Phe}_{CUA}$) was constitutively expressed under control of *lpp* promoter. Th expression cassette of $ytRNA^{Phe}_{CUA}$ was inserted into repressor plasmid pREP4 to form pREP4-

ytRNA as previously described in the Examples herein. The mutant yeast suppressor ytRNA$^{Phe}_{CUA}$_30U40G (ytRNA$^{Phe}_{CUA_UG}$) was constructed from ytRNA$^{Phe}_{CUA}$ by use of a QUICK-CHANGE® mutagenesis kit. Two complementary oligonucleotides, designated as primer UG-f (5'-GAA CAC AGG ACC TCC ACA TTT AGA GTA TGG CGC TCT CCC-3') (SEQ ID NO: 16) for the forward primer and primer UG-r (5'-GGG AGA GCG CCA TAC TCT AAA TGT GGA GGT CCT GTG TTC-3') (SEQ ID NO: 17) for the reverse primer were synthesized to carry the specific mutation at either position 30 or position 40 of mutant yeast suppressor tRNA. The resulting plasmid carrying the gene encoding ytRNA$^{Phe}_{CUA_UG}$ is designated as pREP4-ytRNA_UG. In order to construct the plasmids for *in vitro* transcription of ytRNA^{Phe}, the ytRNA$^{Phe}_{CUA}$ and ytRNA$^{Phe}_{CUA_UG}$ genes were amplified from template plasmid pREP4-ytRNA and pREP4-ytRNA_UG, respectively. At the end of the tRNA sequence, a *BstN*I site was inserted to produce accurate transcript of ytRNA^{Phe}. A T7 promoter sequence was added for *in vitro* transcription of ytRNA^{Phe} by a T7 RNA polymerase. The following primers were used for the PCR: 5'-CTG GGT AAG CTT CGC TAA GGA TCT GCC CTG GTG CGA ACT CTG-3' (SEQ ID NO: 18) (with restriction sites *HindIII and BstNI*) and 5'-GAT TAC GGA TTC CTA ATA CGA CTC ACT ATA GCG GAC TTA GCT C-3' (SEQ ID NO: 19) (with *EcoRI* restriction site and a T7 promoter sequence). The resulting DNA was introduced into the *HindIII* and *EcoRI* sites of pUC18 to yield pUC18-ytRNA$^{Phe}_{CUA}$ and pUC18- ytRNA$^{Phe}_{CUA_UG}$.

In order to facilitate DNA handling, one *BstNI* cleavage site close to the T7 promoter sequence of pUC18- ytRNA$^{Phe}_{CUA}$ was removed to increase the size of the DNA fragment containing the ytRNA$^{Phe}_{CUA}$ gene from 180 bp to 500 bp after *BstNI* digestion. Two complementary oligonucleotides, 5'-CGG AAG CAG AAA GTG TAA AGA GCG GGG TGC CTA ATG AGT G-3' (SEQ ID NO: 20) for the forward primer and 5'-CAC TCA TTA GGC ACC CCG CTC TTT ACA CTT TAT GCT TCC G-3' (SEQ ID NO: 21) for the reverse priemr, were synthesized to carry the specific mutation.

EXAMPLE 12

Linearized DNA was prepared by *BstN*I digestion of pUC18-ytRNA$^{Phe}_{CUA}$ and pUC18-ytRNA$^{Phe}_{CUA_UG}$ as described previously (See Sampson, Uhlenbeck, *PNAS* USA 85: 1033-1037 (1988)). *In vitro* transcription of linearized DNA templates and purification of transcripts were performed as described previously (See Nowak, et al., *Ion Channels* pt. B 293: 504-529 (1998). The *in vitro* transcription of linearized DNA to produce 76mer tRNA transcripts was performed with the AMBION® T7-MEGASHORTSCRIPT® kit. Transcripts were isolated with a 25:24:1 phenol: $CHCl_3$:isoamyl alcohol (PCI) extraction. The organic layer was re-extracted with water and a 24:1 $CHCl_3$:isoamyl alcohol (CI) was performed on the aqueous layers. The water layer was then mixed with an equal volume of isopropanol, precipitated overnight at –20°C, pelleted, dried, and re-dissolved in water. Unreacted nucleotides in the tRNA solution were eliminated using CHROMA SPIN-30® DEPC-H_2O (BD Bioscience®). Concentrations of the transcripts were determined by UV absorbance at 260 nm.

The aminoacylation of wlid-type ytRNA$^{Phe}_{GAA}$ with Phe and Trp by yPheRS variants was performed as described previously (See Sampson, Uhlenbeck, *PNAS* USA 85: 1033-1037 (1988)). Aminoacylation reactions were carried out in the buffer containing 30 mM HEPES (pH 7.45), 15 mM $MgCl_2$, 4mM DTT, 25mM KCl, and 2mM ATP at 30°C, in 100 microliter reaction volumes. Purified yeast total tRNA was used in the assay at final concentration of 4 mg/mL (ytRNA$^{Phe}_{GAA}$ concentration approximately 2.24 microM). For aminoacylation of Phe, 13.3 microM [^{3}H]-Phe (5.3 Ci/mmol) and 80 nM yPheRS variants were used; for aminoacylation of Trp, 3.3 microM [^{3}H]-Trp (30.0 Ci/mmol) and 160 nM yPheRS variants were used. Aminoacylation of ytRNA^{Phe} transcripts was performed in 100 microliter reaction volumes in buffer containing 100 mM potassium-HEPES (pH 7.4), 10 mM $MgCl_2$, 1 mM DTT, 0.2 mM EDTA, 2 mM ATP, and 4 units/mL yeast inorganic pyrophosphatase at 37°C for eLysS. For aminoacylation of Lys, 4 microM of ytRNA^{Phe} transcript,

1.1microM [^{3}H]-Lys (91 Ci/mmol) and 80 nM eLysS were used. The tRNAs were annealed before use by heating up to 85°C for 4 minutes in the annealing buffer (60 mM Tris, pH 7.8, 2 mM $MgCl_2$) followed by slow cooling down to room temperature. Reactions were initiated by adding the enzyme and 10 microliter aliquots were quenched by spotting on Whatman filter disks soaked with 5% TCA. The filters were washed for three 10 minute periods in ice-cold 5% TCA, washed in ice-cold 95% ethanol, and counted via liquid scintillation methods.

EXAMPLE 13

Plasmid construction for *in vivo* incorporation of a non-natural amino acid was performed in a Phe/Trp double auxotrophic strain, AFW, and a Phe/Trp/Lys triple auxotrophic strain, AFWK. The auxotrophic strains were constructed from a Phe auxotrophic strain, AF (K10, Hfr(Cavalli) *pheS13rel-1 tonA22* thi T2^R *pheA*18, *trpB*114) (See Furter, *Prot. Sci*, 7:419-426 (1998)) by P1 phage-mediated transposon transduction. A pQE16 vector (QIAGEN®) was chosen as the expression plasmid, which encodes a marker protein murine dihydrofolate reductase (mDHFR) with C-terminal hexa-histidine tag gene under control of a bacteriophage T5 promoter and t_0 terminator. Quick-change mutagenesis kit was used to place an amber codon (TAG) at the 38th position of mDHFR with two complementary oligonucleotides (5'-CCG CTC AGG AAC GAG TAG AAG TAC TTC CAA AGA ATG-3' (SEQ ID NO: 11); 5'-CAT TCT TTG GAA GTA CTT CTA CTC GTT CCT GAG CGG-3') (SEQ ID NO: 12) to yield pQE16am. The mutant yPheRS genes T415G and T415A were amplified from pQE32-T415G and pQE32-T415A and a constitutive *tac* promoter with an abolished *lac* repressor binding site was added into the upstream of the start codon of the gene. The entire expression cassette of T415G and T415A were inserted into *Pvu*II site of pQE16am-T415G and pQE16am-T415A.

EXAMPLE 14

The auxotrophic bacterial strains AF, AFW, and AFWK were transformed with plasmid pQE16am containing yPheRS variants and pREP4-ytRNA vectors containing ytRNA variants to investigate pBrF incorporation. The *E.coli* expression strains were grown in M9 minimal medium supplemented with glucose, thiamin, $MgSO_4$, $CaCl_2$, 20 amino acids (at 25 mg/L) antibiotics (35 micrograms/mL of kanamycin and 100 micrograms/mL of ampicillin). When cells reached an OD_{600} of 0.8-1.0, they were sedimented by centrifugation, washed twice with cold 0.9% NaCl, and shifted to expression media supplemented with 17 amino acids (at 20 mg/L), 6 mM of pBrF (p-bromo-phenylanine) or plodoF(p-iodo-phenylalanine), and the indicated concentrations of phenylalanine, tryptophan, and lysine. Protein expression was induced by the addition of 1mM IPTG. After four hours expression, cells were pelleted by centrifugation, and the protein was purified by virtue of C-terminal hexa-histidine tag through a nickel-NTA spin column according to manufacturer's directions (QIAGEN®). After purification, expression levels of mDHFR were determined by UV absorbance at 280 nm.

EXAMPLE 15

LC-MS/MS analysis of tryptic digests of mDHFR was conducted on a Finnigan LCQ ion trap mass spectrometer with HPLC pump and ESI probe. Mutant mDHFR purified under denaturing conditions was in elution buffer (8 M urea, 100 mM NaH_2PO_4, 10 mM Tris, pH 4.5). For trypsin digestion, 10 microL of the solution was diluted into 90 microL of 75 mM $(NH_4)_2CO_3$. One microliter of modified trypsin (0.2 micrograms/microliter) was added. The sample was incubated at 37°C for 2 to 6 hours. The digestion reaction was stopped by addition of 12 microL of 5% TFA solution. Digested peptide solution was subjected to desalting with C18 Vydac Microspin column (the Nest group) and eluted with 50 microL of 80% of acetonitrile and 20% of 0.1% w/v formic

acid. Digested peptide solution eluted form Microspin column was dried, redissolved in 10% acetronitrile and 90% of 0.1% TFA solution, and injected into HPLC pump. Peptides were separated by Magic C18 column (Michrom, 300 Å, 0.3 x 150 mm) and eluted at a flow rate of 30 microL/min using a gradient of 10-95% of solvent A (90% of acetonitrile and 10% of 0.1 M acetic acid solution) and solvent B (2% of acetonitrile and 98% of 0.1 M acetic acid solution) for 30 minutes. The column eluent flow to the electrospray source and each signal of tryptic digest was detected. Tandem mass sequencing was carried out simultaneously by fragmentation of the precursor ion with m/z corresponding to protease-digested fragment including the residue at position 38 of mutant mDHFR. Thus, DHFR polypeptides were synthesized in a triple auxotrophic host cell with (a), (b) yeast $tRNA^{Phe}_{CUA}$ and yeast PheRS (T415G); (c) yeast $tRNA^{Phe}_{CUA\_UG}$ and yeast PheRS (T415G); (d) yeast $tRNA^{Phe}_{CUA}$ and yeast PheRS (T415A); (e) yeast $tRNA^{Phe}_{CUA\text{-}UG}$ and yeast PheRS (T415A) or (f) in a single auxotrophic strain with yeast $tRNA^{Phe}_{CUA\_UG}$ and yeast PheRS (T415A), the results of which are shown in Figure 12.

EXAMPLE 16

The binding pocket of TrpRS from *Bacillus sterothermophillus* was mutated in order to incorporate non-natural amino acids into polypeptides. Candidate sites for mutational analysis include amino acid sequence position number 4 (F), 5 (F), 7 (N), 132 (D), 133 (I), 141 (V) and 143 (V) which lie in a region recognized as the hydrophobic amino acid binding pocket.

TrpRS kinetic data for F5Y substitution:

TrpRS	Electrostatics (kcal/mol)	VDW (kcal/mol)	Total
Wild type	-53.94+/- 5.32	-25.78+/-0.35	-79.75+/-5.08
F5Y	-63.04+/-2.47	-25.26+/-0.43	-88.32+/-2.59
Difference	-9.1	+0.5	-8.6

The fused ring of tryptophan gears the recognition site toward the second aromatic ring. As it is more "meta" than "para" in conformation, a mutation of position 132 (D to G) was tested. Briefly, molecular modeling revealed that the 6-ethynyl indole clashes with the Phe5 backbone which inhibited movement. Without backbone movement, the amino acid (analog) will not fit into the binding pocket. As the amino acid at position 132 (D) is highly conserved, we predicted that its modification may disrupt a hydrogen bond network within the TrpRS. Thus, 5-ethynyl tryptophan was computationally modeled in the binding site since it did not clash with the amino acid at location 132. In order to accommodate this analog, the amino acid sequence position 143 was mutated (to A and G, respectively). The binding differentiation was found to be 5.8 kcal/mol (V143A) and 4.4 kcal/mol (V143G) for the binding of 5-ethynyl tryptophan, which distinguish tryptophan from the analog.

Additionally, mutating the amino acid sequence position number 132 to other amino acids was tested. The kinetic data are shown in the table below:

TrpRS kinetic data for mutations in binding site:

Tryptophan	Km (μM)	Kcat (s^{-1})
Wild type	1.6 +/- 0.1	1.1 +/- 0.03
D132N	12.1 +/- 1.6	0.0067 +/- 0.0003
D132S	17.8 +/- 2.3	0.055 +/- 0.004
D132T	8.6+/- 1.4	0.011+/- 0.0008

EXAMPLE 17

A yeast PheRS library (using green fluorescent protein or GFP) was screened to identify PheRS mutations that enable the incorporation of Nal. Specific amino acid sequence positions that were mutated include residue numbers 412, 415, 418, and 437, which are located in the binding site and

contact the amino acid. As indicated in Figure 13, *p*-ethynyl-phenylalanine was incorporated into a test protein using the modified yeast PheRS.

Briefly, GFP was ligated into a vector containing the mutant T415G or wild type yeast PheRS gene according to standard procedures. The mutant yeast amber suppressor tRNA ($ytRNA^{Phe}_{AAA}$) was constitutively expressed under control of *lpp* promoter and transformed into a Phe/Trp double *E.coli* auxotrophic strain AFW. Cells were grown in M9 minimal medium supplemented with glucose, thiamin, $MgSO_4$, $CaCl_2$, 20 amino acids (at 25 mg/L), antibiotics (35 µg/mL of kanamycin and 100 µg/mL of ampicillin). When cells reached an OD_{600} of 0.8-1.0, cells were pelleted and washed twice by ice-cold 0.9% NaCl and shifted to expression media supplemented with 18 amino acids (at 20 mg/L) and various concentrations of phenylalanine, tryptophan and 2Nal. Protein expression was induced by IPTG.

EXAMPLE 18

Mutagenesis of the four amino acid residues selected (N412, T145, S418, and S437) were conducted by two step PCR mutation. Briefly, a series of PCR mutagenesis were performed at GFP_{UV} gene in a pQE9_GFP_{UV} plasmid (STRATAGENE®), using four complementary pairs of primers (F64LS65T_f: 5'-CTT GTC ACT ACT CTG ACC TAT GGT GTT CAA TGC TTC TCC CGT-3' (SEQ ID NO: 22); F64LS65T_r : 5'-ACG GGA GAA GCA TTG AAC ACC ATA GGT CAG AGT AGT GAC AAG-3' (SEQ IDNO: 23); S99F_f: 5'-GTA CAG GAA CGC ACT ATA TTC TTC AAA GAT GAC GGG AAC-3' (SEQ ID NO: 24); S99F_r : 5'-GTT CCC GTC ATC TTT GAA GAA TAT AGT GCG TTC CTG TAC-3' (SEQ ID NO: 25); T153M_f : 5'- CAC AAT GTA TAC ATC ATG GCA GAC AAA CAA AAG AAT GGA-3' (SEQ ID NO: 26); T153M_r : 5'-TCC ATT CTT TTG TTT GTC TGC CAT GAT GTA TAC ATT GTG -3' (SEQ ID NO: 27)).

The GFP mutants were generated as described herein. Briefly, a GFP3 has 12 Phe residues of which five are encoded by Phe wobble codons

(UUU). A GFP5 and a GFP6 variant were prepared by replacing UUC codons with UUU codons using two-step PCR reactions followed by ligation. A GFP5 was prepared by replacing four UUC codons and one Leu codon at F8, L64, F84, F99 and F165 residues with UUU codons using twelve primers (1: 5'-GTG CCA CCT GAC GTC TAA GAA ACC ATT ATT ATC ATG ACA TTA ACC-3' (SEQ ID NO: 28) 2: 5'-GAG TAA AGG AGA AGA ACT TTT TAC TGG AGT TGT CCC AAT TC-3' (SEQ ID NO: 29) 3: 5'- GAA TTG GGA CAA CTC CAG TAA AAA GTT CTT CTC CTT TAC TC-3' (SEQ ID NO: 30) 4: 5'- GGC CAA CAC TTG TCA CTA CTT TTA CCT ATG GTG TTC AAT GCT T-3' (SEQ ID NO: 31) 5: 5'- AAG CAT TGA ACA CCA TAG GTA AAA GTA GTG ACA AGT GTT GGC C-3' (SEQ ID NO: 32) 6:5'- CAT ATG AAA CGG CAT GAC TTT TTT AAG AGT GCC ATG CCC GAA G-3' (SEQ ID NO:33) 7: 5'- CTT CGG GCA TGG CAC TCT TAA AAA AGT CAT GCC GTT TCA TAT G (SEQ ID NO: 34) 8: 5'- GTT ATG TAC AGG AAC GCA CTA TAT TTT TCA AAG ATG ACG GGA ACT ACA A-3' (SEQ ID NO: 35) 9:5'- TTG TAG TTC CCG TCA TCT TTG AAA AAT ATA GTG CGT TCC TGT ACA TAA C-3' (SEQ ID NO: 36) 10: 5'- ACA AAA GAA TGG AAT CAA AGC TAA CTT TAA AAT TCG CCA CAA CAT TGA AGA TG-3' (SEQ ID NO: 37) 11: 5'- CAT CTT CAA TGT TGT GGC GAA TTT TAA AGT TAG CTT TGA TTC CAT TCT TTT GT-3' (SEQ ID NO: 38); 12: 5'- CGC CAA GCT AGC TTG GAT TCT CAC CAA TAA AAA ACG CCC-3' (SEQ ID NO: 39) Five partially overlapping fragments of GFP3 expression cassettes were obtained by five PCR reactions with five sets of primers (1 and 3; 2 and 5; 4 and 7; 6 and 9; 8 and 10).

These PCR products were purified by agarose gel electrophoresis followed by gel extraction. A GFP6 of which all Phe residues are encoded by UUU was prepared by replacing two Phe codons (F71 and F99) of GFP5 with UUU codons using six primers. (1 and 12 are the same as above; 13: 5'- TAC CTA TGG TGT TCA ATG CTT TTC CCG TTA TCC GGA TCA TAT G-3' (SEQ ID NO: 40); 14: 5'-CAT ATG ATC CGG ATA ACG GGA AAA GCA TTG AAC ACC ATA GGT A-3' (SEQ ID NO: 41); 15: 5'- GTT ATG TAC AGG AAC GCA

CTA TAT TTT TTA AAG ATG ACG GGA ACT ACA AG-3' (SEQ ID NO: 42); 16: 5'- CTT GTA GTT CCC GTC ATC TTT AAA AAA TAT AGT GCG TTC CTG TAC ATA AC-3' (SEQ ID NO: 43)).

EXAMPLE 19

Library construction was performed in two steps as well. Briefly, saturation mutagenesis in four residues (N412, T415, S418 and S437) was accomplished with two step PCR mutagenesis. First, degenerate codons were introduced into S437 by PCR mutagenesis with two complementary primers (437_f: 5'-GTC GAA ATC GGT AAC NNK GGT ATG TTC AGA CCA GAA ATG CTC G-3' (SEQ ID NO: 44); 437_r: 5'- C GAG CAT TTC TGG TCT GAA CAT ACC MNN GTT AC C GAT TTC GAC-3' (SEQ ID NO: 45)). After 1 hr digestion of PCR product with *Dpn*I, PCR product was transformed into XL-1 blue cloning host. The plasmids of the 437th position were saturated and isolated and used as a template for 2nd PCR mutagenesis to introduce mutation at residues N412, T415 and S418. The 2nd PCR mutagenesis was performed with another complementary primer pair (412_418_f: 5'-C AAG CCT ACC TAC NNK CCT TAC NNK GAG CCA NNK ATG GAA ATC TTT T-3' (SEQ ID NO: 46); 412_418_r: 5'- A AAA GAT TTC CAT MNN TGG CTC MNN GTA AGG MN N GTA GG T AGG CTT G-3' (SEQ ID NO: 47)). Following PCR, the products were digested with *Dpn*I for 1 hr, it was cleaned and concentrated by spin column. Elute was electroporated into ElectroTen –Blue electrocompetent cell (Stratagene) according to manufacturer's protocol. Eight million transformants were obtained. The library plasmid was expanded in culture and digested with *Nsi*I and *Bgl*II. After purification of these inserts, they was ligated with large fragments of pQE9_GFP6_yPheRS (T415G) and pQE9_GFP9_yPheRS (T415G) obtained by digestion with *Nsi*I and *Bgl*II.

The library was transformed into chemical-competent AFW and DHF *E.coli* cells. These cells were then inoculated into 2x YT media with kanamycin and grown overnight. When cells reached an OD_{600} of 0.8, cells

were pelleted and resuspended in distilled water. Glycerol stocks of the library were expressed as is standard in the art.

After expression of GFP for 3 hours, 1 mL of cells (based on OD_{600} of 1.0) were washed with PBS and diluted in distilled water, then subjected to flow cytometric analysis (MoFlo cell sorter®, DakoCytomation, Fort Collins, CO), using an excitation wavelength of 488 nm, emission of 525 nm, and a cut-off filter of 495 nm. At least 20,000 events were collected in each measurement. Data were analyzed with Summit software (DakoCytomation). Library screening was done both positively and negatively, that is the yPheRS variants that enable the high incorporation of 2Nal or any other natural amino acids except Phe at UUU codons will unfold GFP and are less bright, and so low fluorescence cells are collected. The yPheRS variants that do not allow incorporation of any other natural amino acids except Phe at UUU codons will not affect GFP folding and are bright. Thus, bright cells are collected. Figure 14 illustrates histograms of GFP yPheRS library screening.

EXAMPLE 20

A modified MetRS from *E.coli* that was mutated at amino acid sequence position 13 (L→G) to incorporate azidonorleucine into a test protein (DHFR) in plasmid pQE-80, according to the methods described herein for other Examples, and at SEQ ID NO:1. In this particular exemplary embodiment, the DHFR and MetRS genes are located in the same plasmid vector.

EXAMPLE 21

Interferon-beta molecule was used as a test molecule to mutate three out of four methionine residues to other replacement amino acids (including non-natural amino acids). Methionine residues at amino acid positions 36, 62 and 117 were mutated to other amino acids via side chain

rotamer excitation analysis. Structures were optimized using molecular dynamics software and the energy calculations of the mutated structures, including salvation. Next, comparisons were made of the energy calculations of the wild type interferon beta molecule with the modified interferon beta molecule in order to determine the overall stability of the modified molecule. Results of energy calculations with the various point mutations are shown in the tables below:

Mutation at Position 36 of human interferon beta:

Mutation	Energy
M→H	-0.2 kcal/mol
M→C	-0.2 kcal/mol
M→I	+1.0 kcal/mol
M→T	+1.4 kcal/mol
M→V	+1.6 kcal/mol
M→A	+4 kcal/mol

Mutation at Position 62 of human interferon beta:

Mutant	Energy (kcal/mol)
H	+1.1
G	0
Y	-2.2
S	-4.7
Q	-4.8
A	-5.0
N	-5.7
F	-7.4
T	-8.8
Wild type	-11.6

Mutation at Position 117 of human interferon beta:

Mutation	Energy
M→I	-1.2 kcal/mol
M→L	-1.0 kcal/mol
M→V	-0.1 kcal/mol
M→T	+3 kcal/mol
M→Y	+3 kcal/mol
M→S	+4 kcal/mol
M→G	+5.9 kcal/mol

All of the above U.S. patents, U.S. patent application publications, U.S. patent applications, foreign patents, foreign patent applications, and non-patent publications referred to in this specification and/or listed in the Application Data Sheet, are herein incorporated by reference in their entireties.

From the foregoing it will be appreciated that, although specific embodiments of the invention have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of the invention. Accordingly, the invention is not limited except as by the appended claims.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, numerous equivalents to the specific method and reagents described herein, including alternatives, variants, additions, deletions, modifications and substitutions. Such equivalents are considered to be within the scope of this invention and are covered by the following claims.

CLAIMS

1. A composition comprising a first vector containing a polynucleotide encoding a modified amino-acyl tRNA synthetase (AARS), wherein said polynucleotide modified synthetase is mutated at one or more codons encoding the amino acid binding region necessary for interaction with the amino acid to be paired with a tRNA molecule, and wherein said modified synthetase is capable of charging a tRNA molecule with a non-natural amino acid.

2. The composition of claim 1, wherein said binding region comprises no more than 30, 20, 15, 10, or 5 contiguous amino acid residues.

3. The composition of claim 1, wherein said modified AARS is selected from the group consisting of a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS.

4. The composition of claim 3 wherein said PheRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 412, 415, 418, and 437.

5. The composition of claim 3 wherein said TrpRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 4, 5, 7, 132, 133, 141, and 143.

6. The composition of claim 3 wherein said MetRS is mutated at amino acid sequence position number 13.

7. The composition of claim 1, further comprising a second vector containing a polynucleotide encoding a tRNA molecule.

8. The composition of claim 7 wherein said first and second vectors are the same vector.

9. The composition of claim 7 wherein said first and second vectors are different vectors.

10. The composition of claim 7 wherein said tRNA is modified.

11. The composition of claim 10 wherein said tRNA is modified such that it contains a mutated anticodon that base pairs with a corresponding wobble degenerate codon with an affinity greater than the affinity of the natural tRNA.

12. The composition of claim 1, wherein said AARS and said tRNA are from the same or different organisms.

13. The composition of claim 1 wherein said non-natural amino acid is selected from the group consisting of: azidonorleucine, 3-(1-naphthyl)alanine, 3-(2-naphthyl)alanine, *p*-ethynyl-phenylalanine, *p*-propargly-oxy-phenylalanine, *m*-ethynyl-phenylalanine, 6-ethynyl-tryptophan, 5-ethynyl-troptophan, (R)-2-amino-3-(4-ethynyl-1H-pyrol-3-yl)propanic acid, *p*-bromophenylalanine, *p*-idiophenylalanine, p-azidophenylalanine, 3-(6-chloroindolyl)alanine, 3-(6-bromoindolyl)alanine, 3-(5-bromoindolyl)alanine, azidohomoalanine, and *p*-chlorophenylalanine.

14. A polypeptide comprising a modified amino-acyl tRNA synthetase (AARS), wherein said modified synthetase is mutated at one or more codons in the amino acid binding region necessary for interaction with the amino acid to be paired with a tRNA molecule, and wherein said modified

synthetase is capable of charging a tRNA molecule with a non-natural amino acid.

15. The polypeptide of claim 14, wherein said binding region comprises no more than 30, 20, 15, 10, or 5 contiguous amino acid residues.

16. The polypeptide of claim 14, wherein said modified AARS is selected from the group consisting of a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS.

17. The polypeptide of claim 16 wherein said PheRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 412, 415, 418, and 437.

18. The polypeptide of claim 16, wherein said TrpRS is mutated at amino acid sequence positions selected from the group consisting of amino acid sequence position number 4, 5, 7, 132, 133, 141, and 143.

19. The polypeptide of claim 16, wherein said MetRS is mutated at amino acid sequence position number 13.

20. A translation system comprising the polynucleotide of claim 1.

21. The translation system of claim 20 wherein said system comprises a host cell.

22. The translation system of claim 21 wherein said modified amino-acyl tRNA synthetase is derived from an organism different than the host cell.

23. The translation system of claim 20 further comprising a polynucleotide encoding a modified tRNA molecule.

24. The translation system of claim 23 wherein said modified tRNA molecule is derived from an organism different than the host cell.

25. The translation system of claim 23 wherein said modified tRNA molecule is derived from a eukaryotic cell and the host cell is a prokaryotic cell.

26. The host cell of claim 21 wherein the cell is an auxotroph.

27. The translation system of claim 20 further comprising a culture media containing one or more non-natural amino acids.

28. The translation system of claim 20 wherein said one or more non-natural amino acids are selected from the group consisting of: azidonorleucine, 3-(1-naphthyl)alanine, 3-(2-naphthyl)alanine, *p*-ethynyl-phenylalanine, *p*-propargly-oxy-phenylalanine, *m*-ethynyl-phenylalanine, 6-ethynyl-tryptophan, 5-ethynyl-troptophan, (R)-2-amino-3-(4-ethynyl-1H-pyrol-3-yl)propanic acid, *p*-bromophenylalanine, *p*-idiophenylalanine, p-azidophenylalanine, 3-(6-chloroindolyl)alanine, 3-(6-bromoindolyl)alanine, 3-(5-bromoindolyl)alanine, azidohomoalanine, and *p*-chlorophenylalanine.

29. The translation system of claim 20 wherein said modified AARS is selected from the group consisting of: a modified PheRS, a modified TrpRS, a modified TyrRS, and a modified MetRS.

30. A method for incorporating a non-natural amino acid into a target polypeptide at one or more specified position(s), the method comprising the steps of:

(1) determining the structural change in the polypeptide for incorporation of a non-natural at one specific position in the polypeptide;

(2) providing a translation system;

(3) providing to the translation system a first polynucleotide of claim 1, or the modified AARS encoded thereby;

(4) providing to the translation system the non-natural amino acid;

(5) providing to the translation system a template polynucleotide encoding a polypeptide of interest, and,

(6) allowing translation of the template polynucleotide, thereby incorporating the non-natural amino acid into the polypeptide of interest at the specified position(s),

wherein steps (1)-(4) are effectuated in any order.

31. The method of claim 30, wherein said translation system comprises a cell.

32. The method of claim 30, wherein step (4) is effectuated by contacting said translation system with a solution containing the non-natural amino acid.

33. The method of claim 30, wherein the specificity constant (k_{cat} / K_M) for activation of said non-natural amino acid by said modified AARS is at least 5-fold larger than that for said natural amino acid.

34. The method of claim 30, wherein said modified AARS mischarges a tRNA at a rate of no more than 1%, 2%, 3%, 4%, 5%, 6%, 7%, or 8%.

35. The method of claim 34, wherein said tRNA is a modified tRNA.

36. The method of claim 35, wherein said first polynucleotide or said second polynucleotide further comprises either a constitutively active or an inducible promoter sequence that controls the expression of the tRNA or AARS.

37. The method of claim 30 further comprising the step of screening for cells containing a modified AARS.

38. The method of claim 30, further comprising the step of verifying the incorporation of the non-natural amino acid.

39. The method of claim 30 wherein said modified AARS is selected from the group consisting of: PheRS, TyrRS, TrpRS, and MetRS.

40. A polypeptide made by the method of claim 30.

41. A method for incorporating at least one non-natural amino acid into a target polypeptide at one or more specified location(s), the method comprising providing a translation system containing at least one non-natural amino acid; providing to the translation system one or more modified AARS selected from the group consisting of: modified PheRS, TrpRS, TyrRS, and MetRS; providing to the translation system a polynucleotide encoding a target

polypeptide of interest; and allowing translation of interest, thereby incorporating at least one non-natural amino acid into the target polypeptide.

42. A polypeptide made by the method of claim 41.

	261	314
T. thermophilus:	RFQPVYFPFVEP	...GFAFGLGVERLAMLRY
E. coli:	RFRPSTFPFTEP	...GFAFGMGMERLTMLRY
S. cerevisiae:	RFKPTYNPYTEP	...VLGMGLSLERPTMIKY
	415	460

FIG. 1A

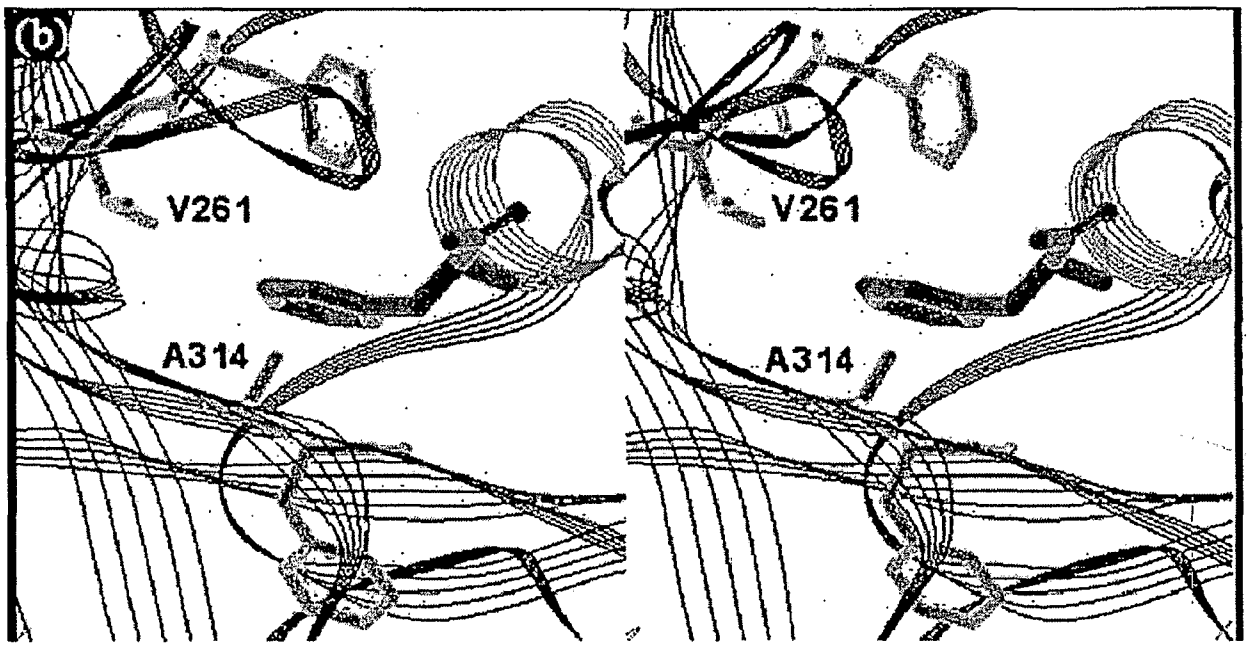


FIG. 1B

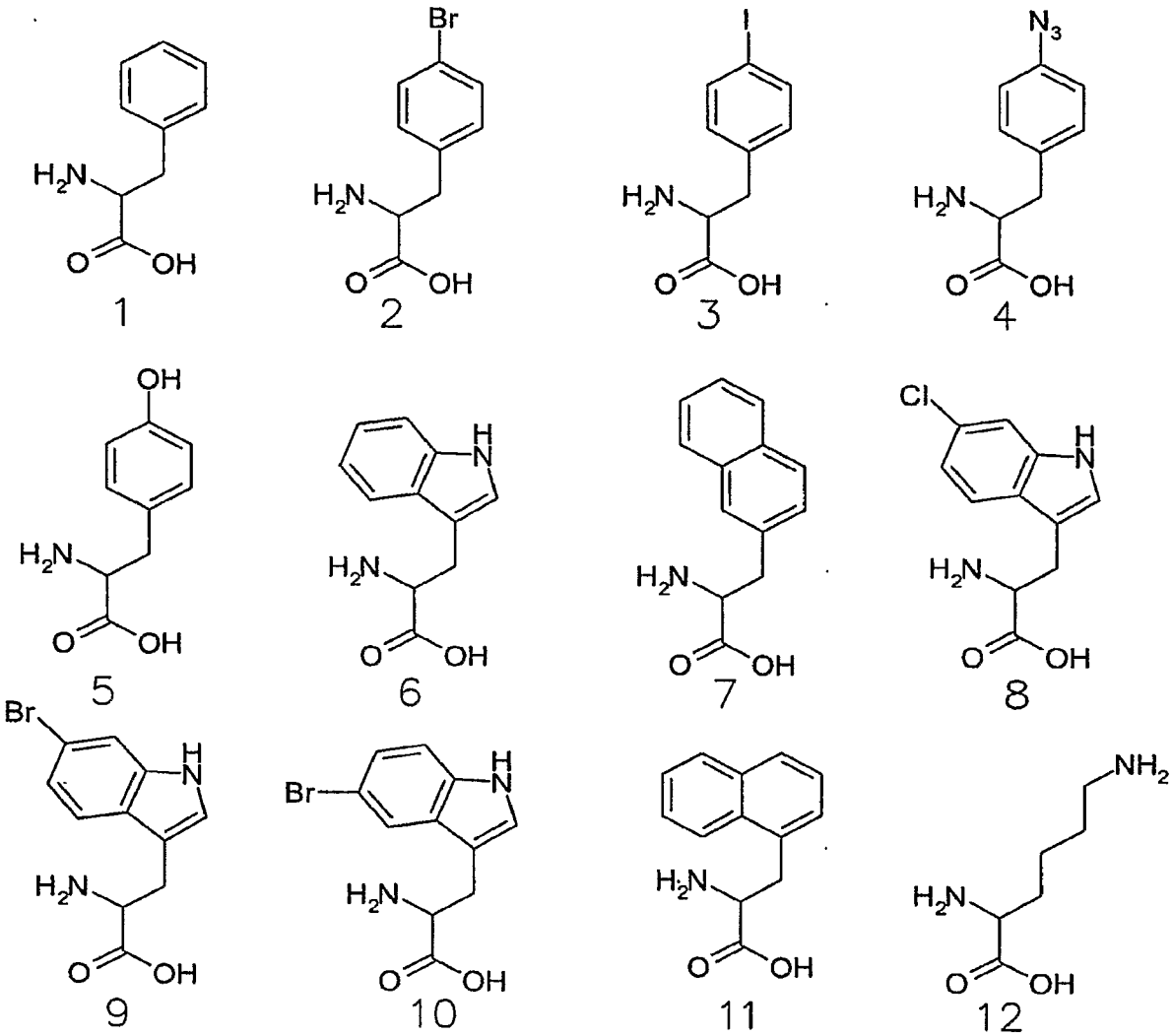


FIG. 2

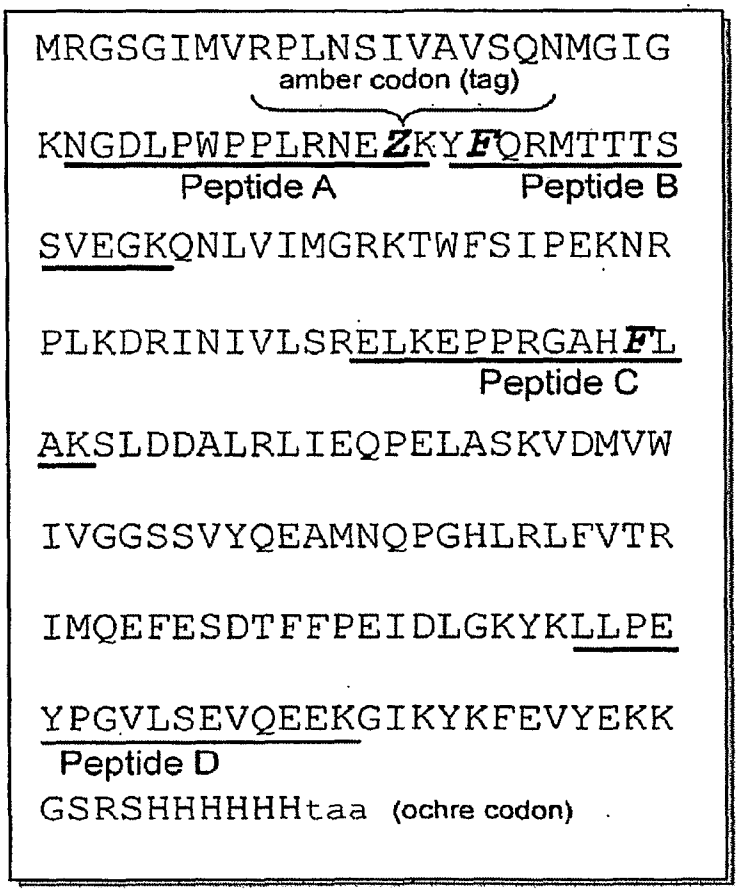

MRGSGIMVRPLNSIVAVSQNMGIG
amber codon (tag)
KNGDLPWPPLRNEZKYFQRMTTTS
Peptide A
Peptide B
SVEGKQNLVIMGRKTWFSIPEKNR
PLKDRINIVLSRELKEPPRGAHFL
Peptide C
AKSLDDALRLIEQPELASKVDMVW
IVGGSSVYQEAMNQPGHLRLFVTR
IMQEFESDTFFPEIDLGKYKLLPE
YPGVLSEVQEEKGIKYKFEVYEKK
Peptide D
GSRSHHHHHHtaa (ochre codon)

FIG. 3

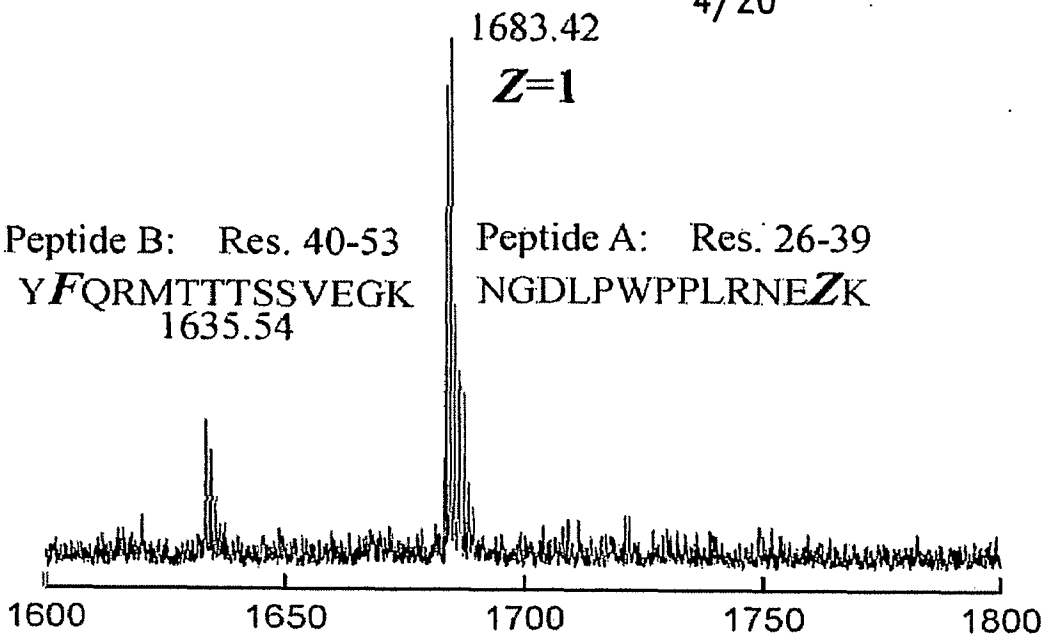

1683.42
Z=1
Peptide B: Res. 40-53
YFQRMTTTSSVEGK
1635.54
Peptide A: Res. 26-39
NGDLPWPPLRNEZK
1600
1650
1700
1750
1800

FIG. 4A

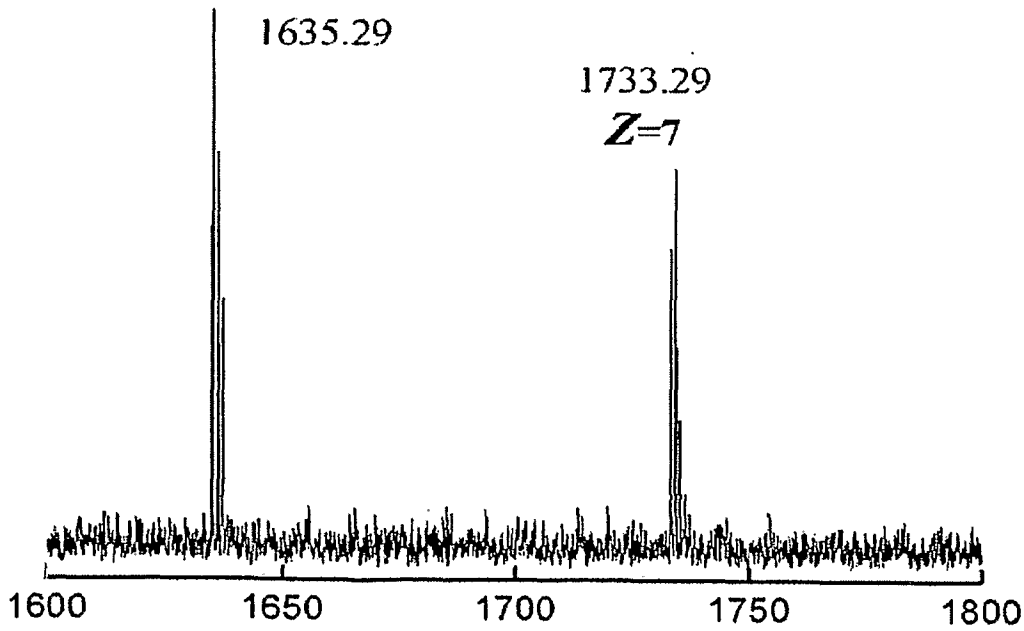

1635.29
1733.29
Z=7
1600
1650
1700
1750
1800

FIG. 4B

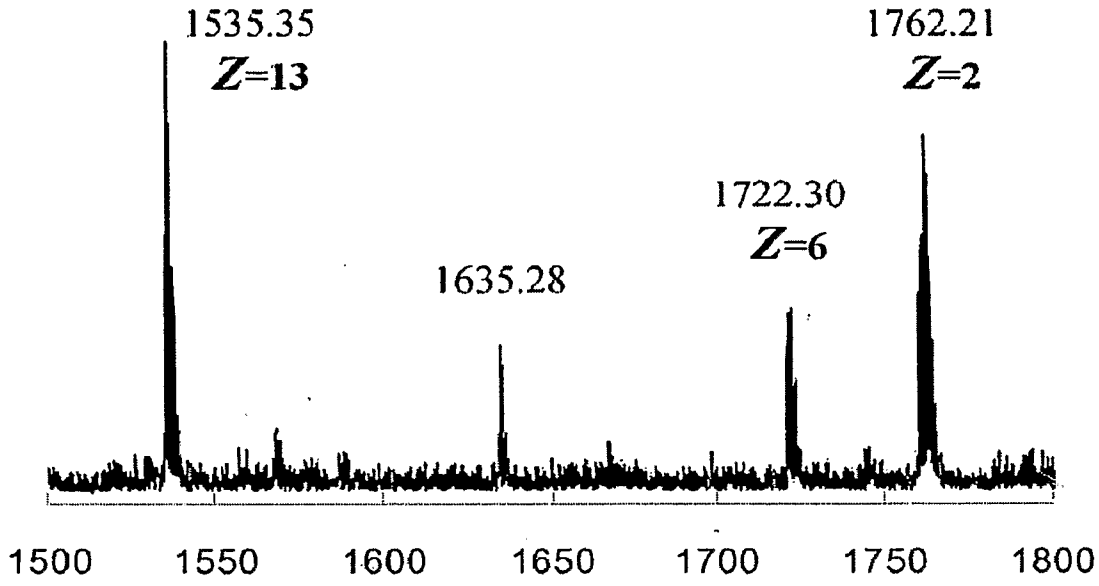

1535.35
Z=13
1762.21
Z=2
1722.30
Z=6
1635.28
1500
1550
1600
1650
1700
1750
1800

FIG. 4C

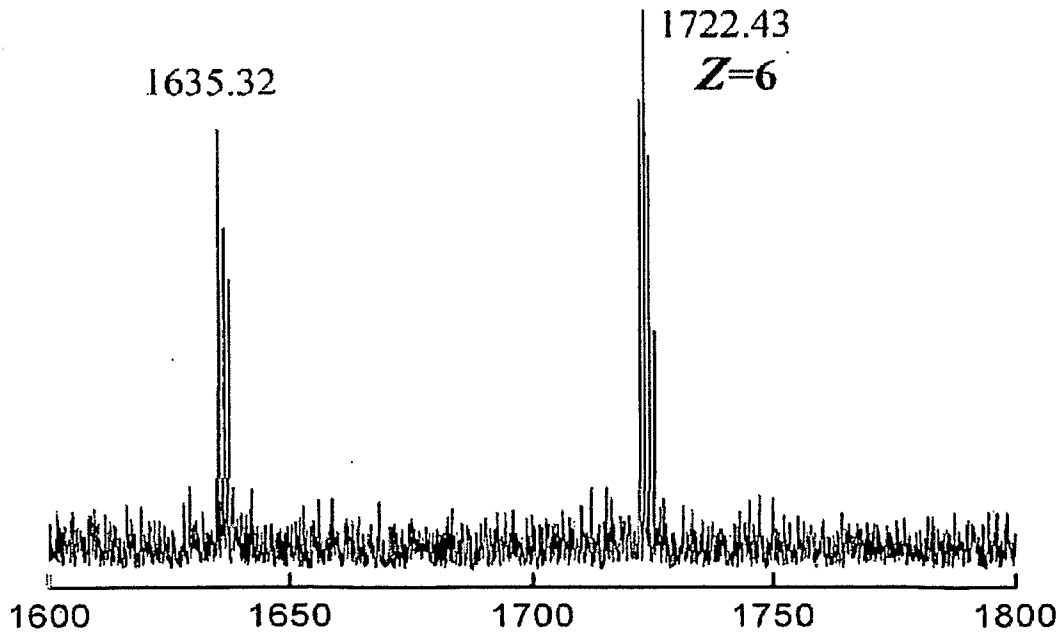

1722.43
Z=6
1635.32
1600
1650
1700
1750
1800

FIG. 4D

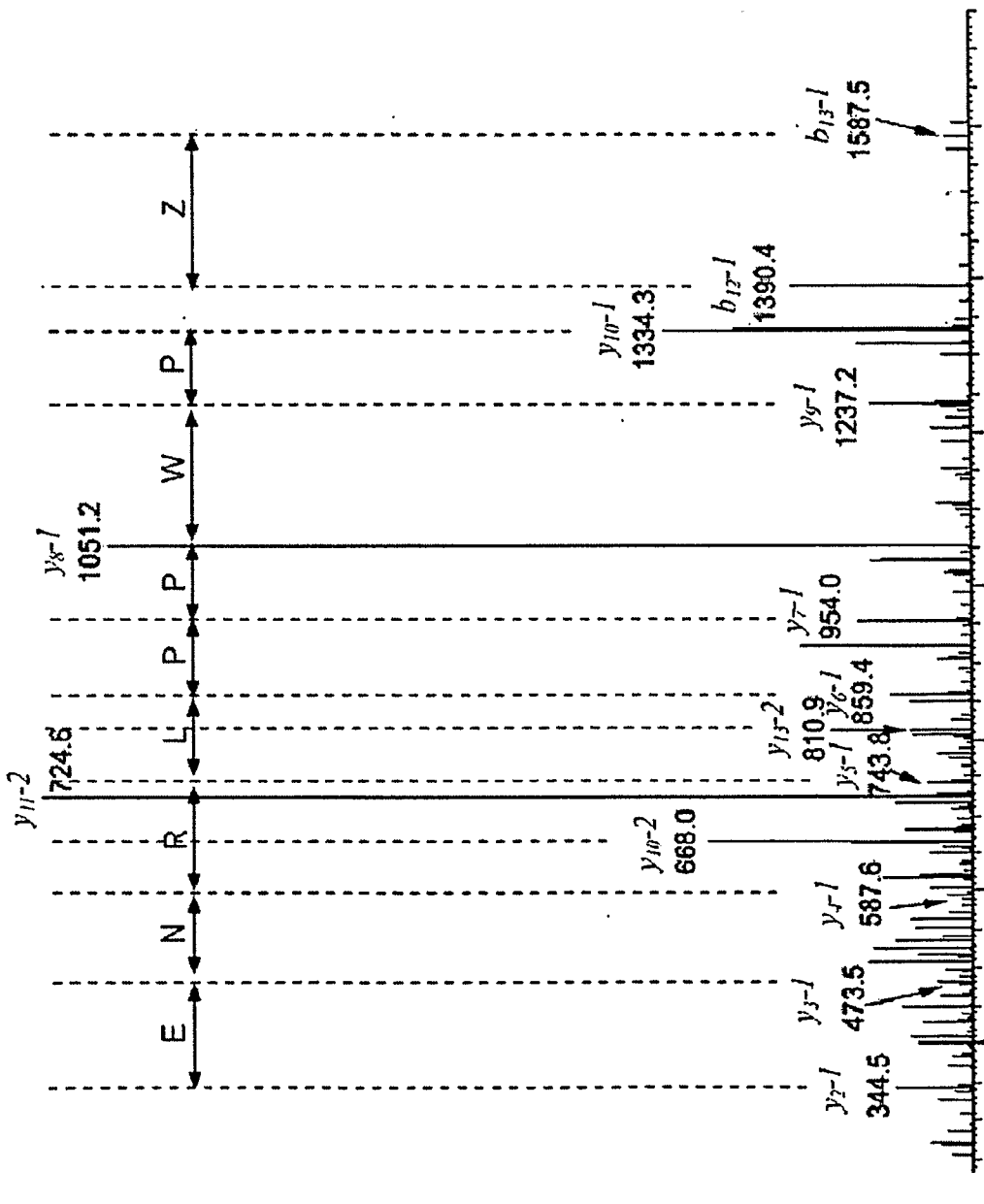

E
N
R
L
P
P
W
P
Z
y_{11}-2
724.6
y_8-1
1051.2
y_{10}-2
668.0
y_{10}-1
1334.3
y_{13}-2
810.9
y_{13}-1
1390.4
y_7-1
954.0
y_5-1
743.8
y_6-1
859.4
y_9-1
1237.2
b_{13}-1
1587.5
y_4-1
587.6
y_3-1
473.5
y_2-1
344.5
200
400
600
800
1000
1200
1400
1600
m/z

FIG. 5

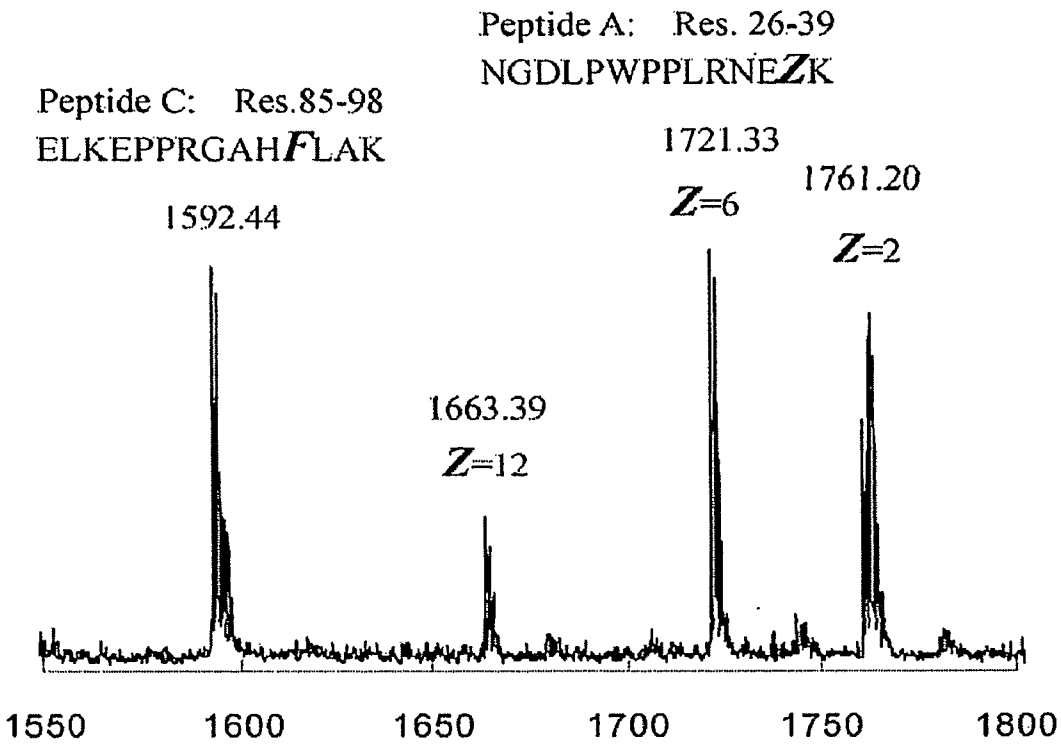

Peptide A: Res. 26-39
NGDLPWPPLRNEZK
Peptide C: Res.85-98
ELKEPPRGAHFLAK
1592.44
1663.39
Z=12
1721.33
Z=6
1761.20
Z=2
1550
1600
1650
1700
1750
1800

FIG. 6A

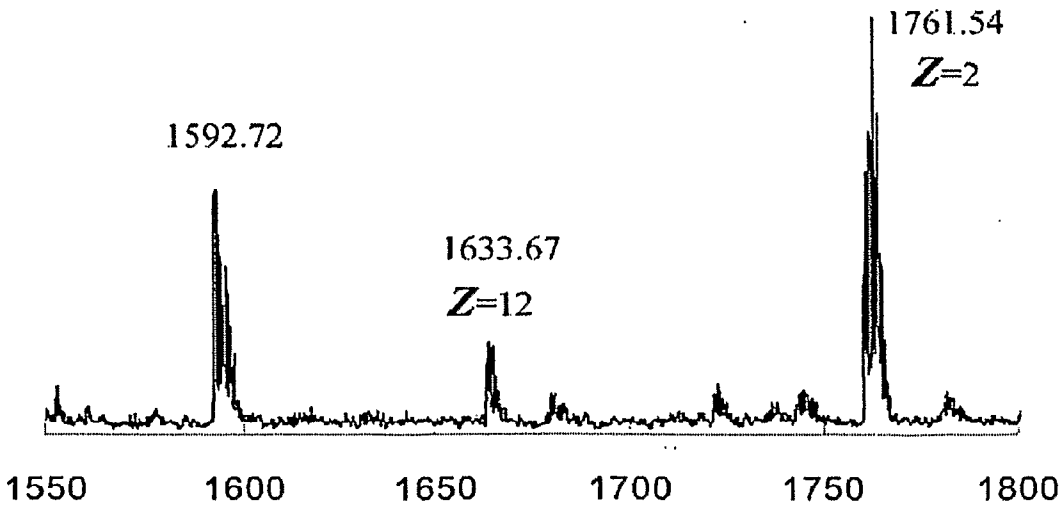

1592.72
1633.67
Z=12
1761.54
Z=2
1550
1600
1650
1700
1750
1800

FIG. 6B

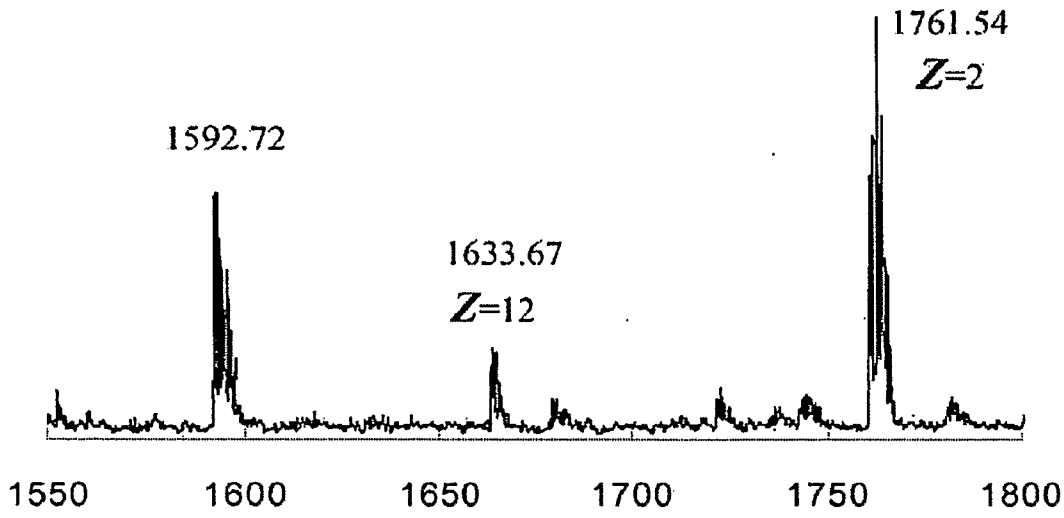

1761.54
Z=2
1592.72
1633.67
Z=12
1550
1600
1650
1700
1750
1800

FIG. 6C

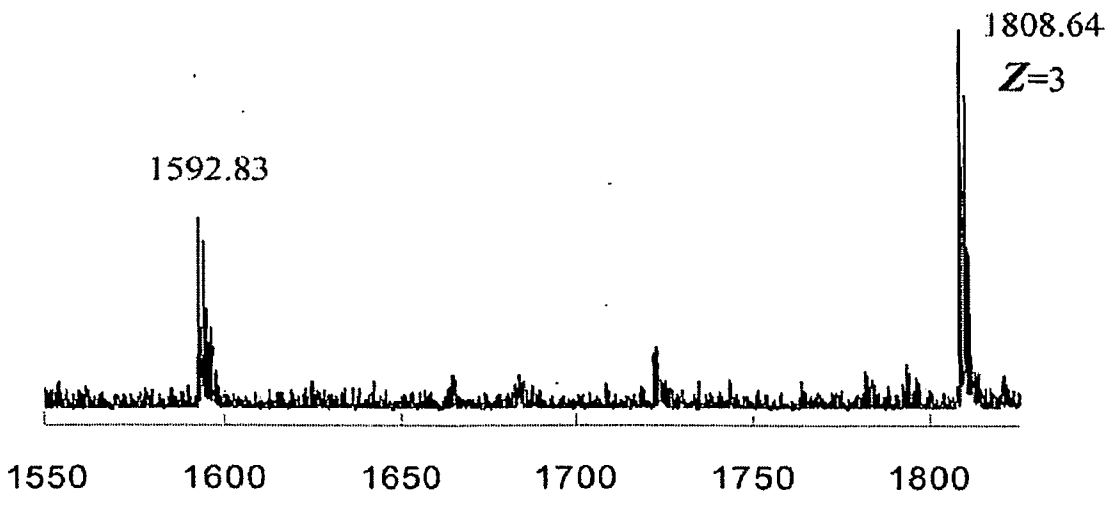

1808.64
Z=3
1592.83
1550
1600
1650
1700
1750
1800

FIG. 6D

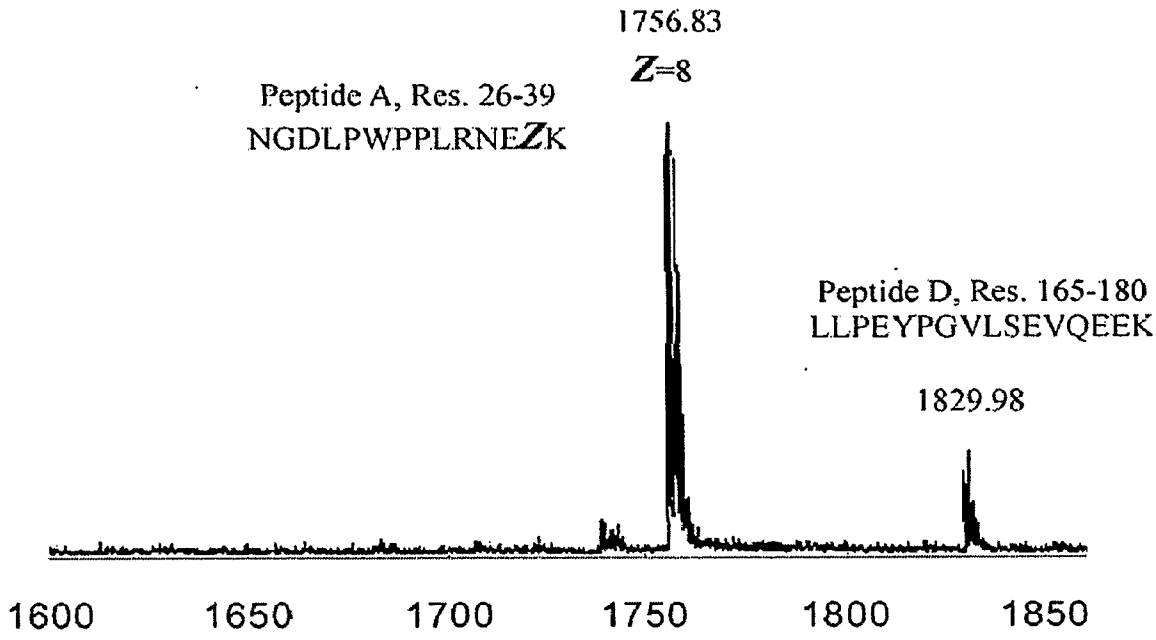

1756.83
Z=8
Peptide A, Res. 26-39
NGDLPWPPLRNEZK
Peptide D, Res. 165-180
LLPEYPGVLSEVQEEK
1829.98
1600
1650
1700
1750
1800
1850

FIG. 7A

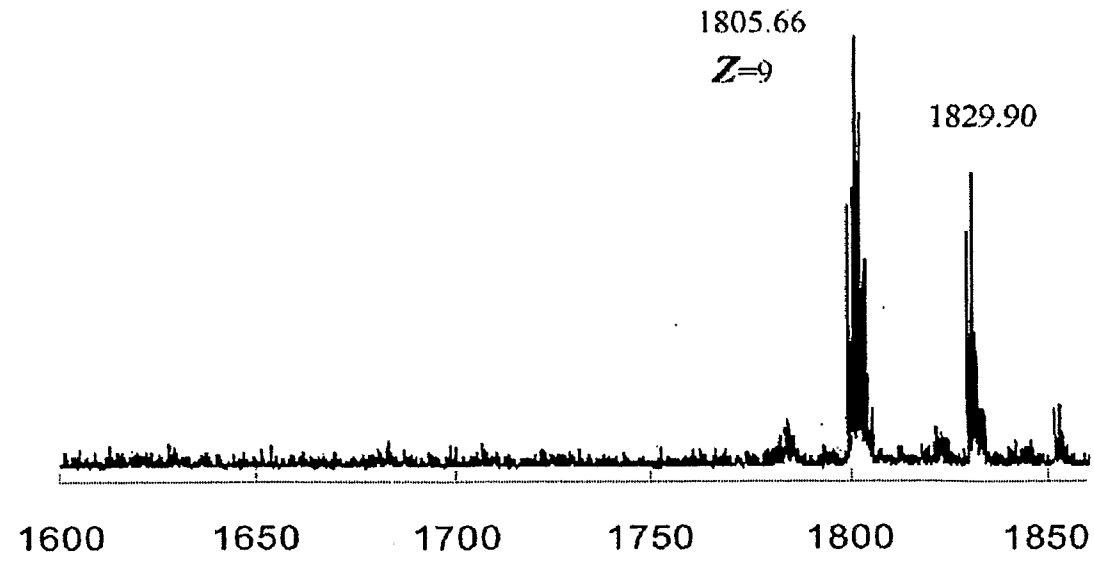

1805.66
Z=9
1829.90
1600
1650
1700
1750
1800
1850

FIG. 7B

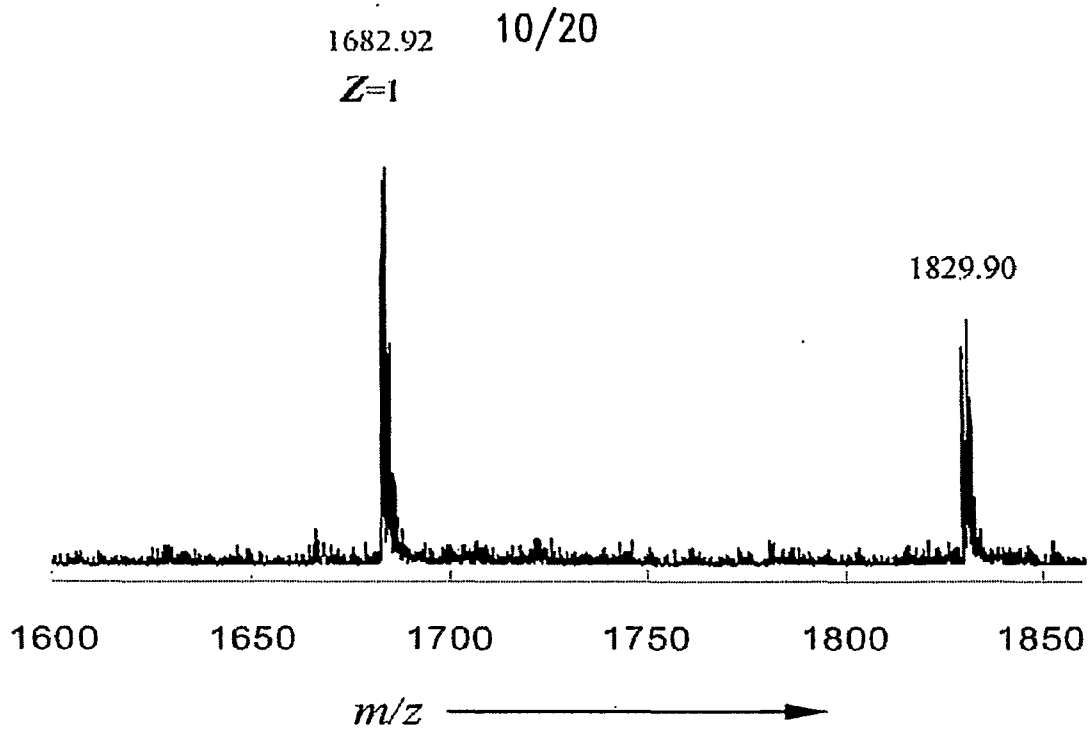

1682.92
Z=1
1829.90
1600
1650
1700
1750
1800
1850
m/z

FIG. 7C

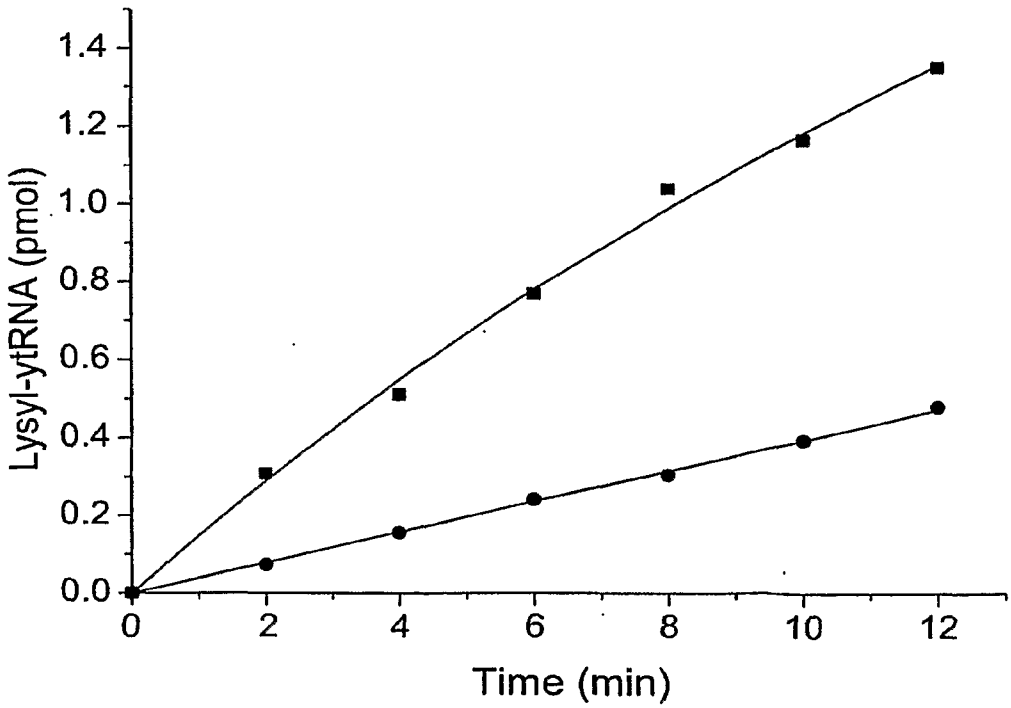

Lysyl-ytRNA (pmol)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
Time (min)

FIG. 8

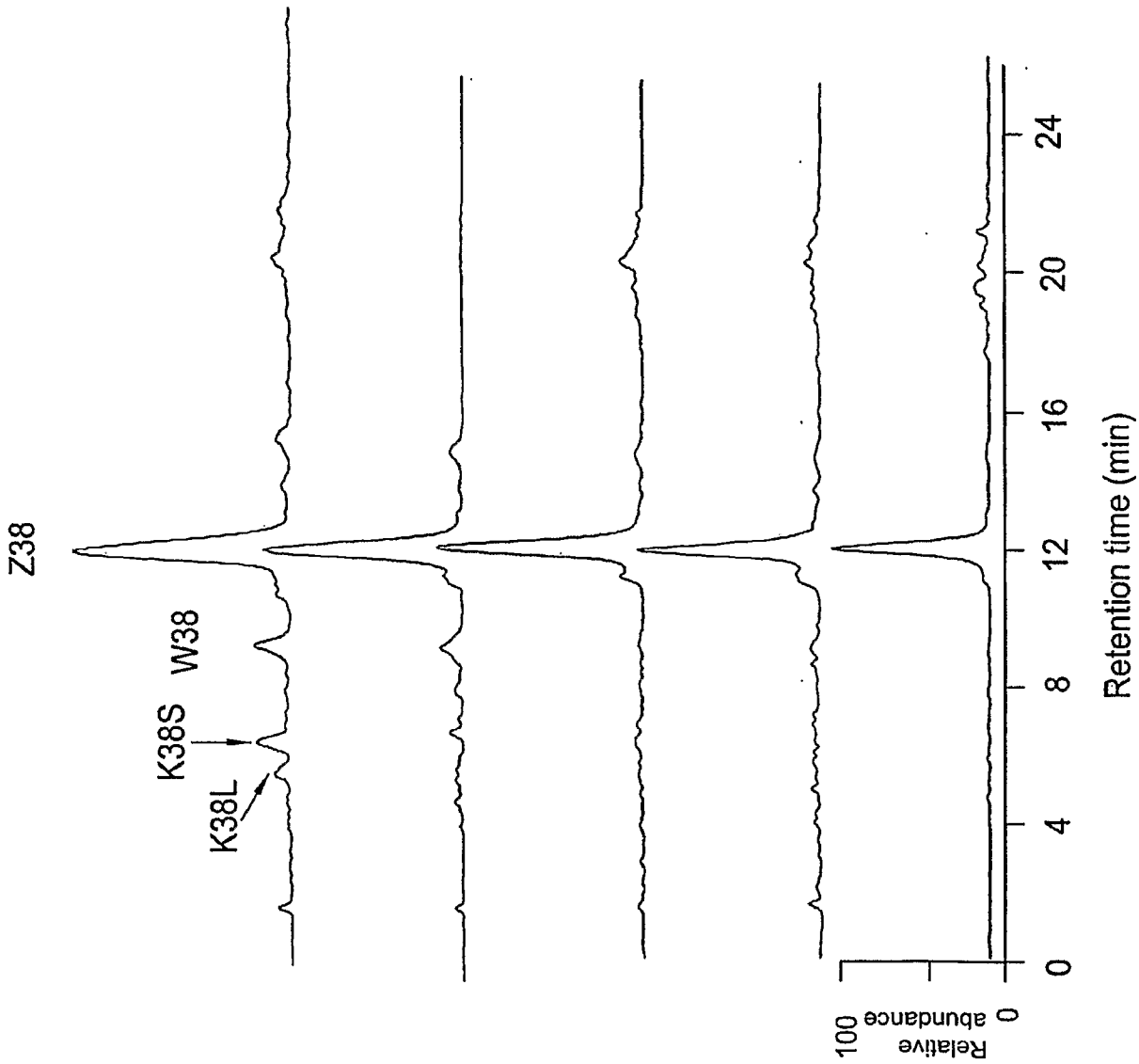


FIG. 9A

FIG. 9B

FIG. 9C

FIG. 9D

FIG. 9E

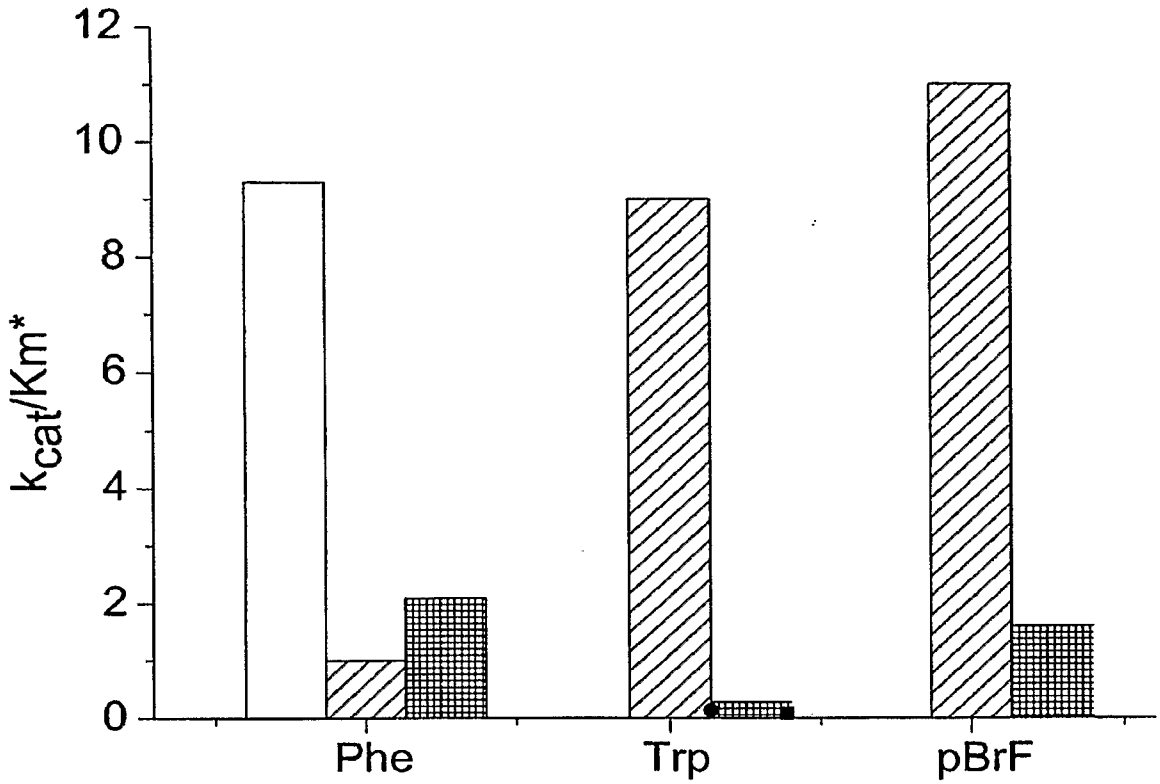


FIG. 10

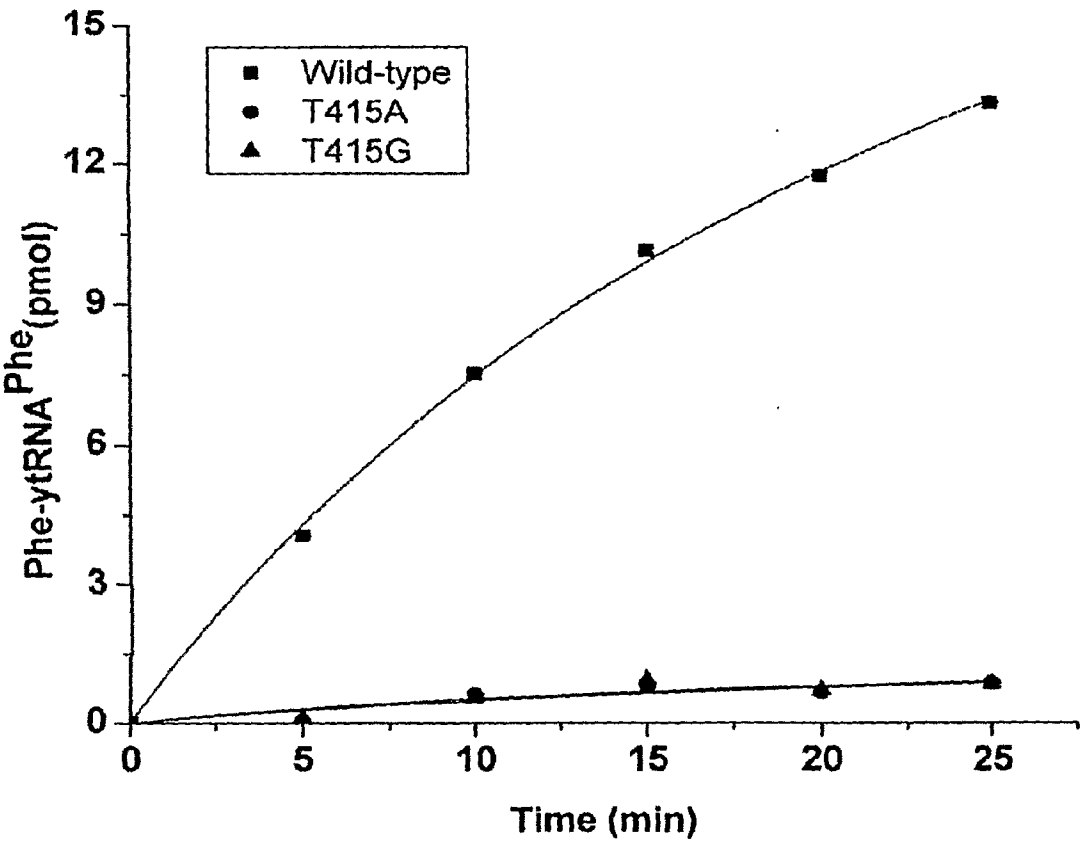

15
12
9
6
3
0
0
5
10
15
20
25
Wild-type
T415A
T415G
Phe-ytRNA^Phe(pmol)
Time (min)

FIG. 11A

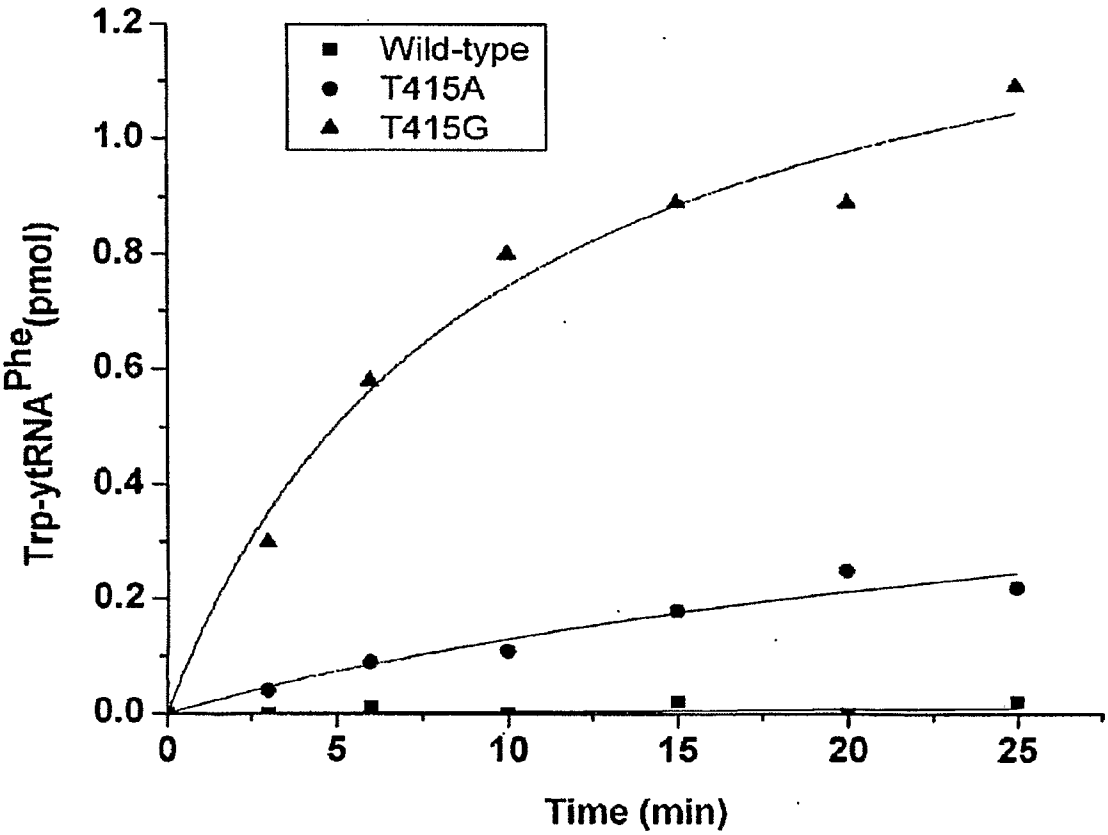

Wild-type
T415A
T415G
Trp-ytRNA^{Phe}(pmol)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
Time (min)

FIG. 11B

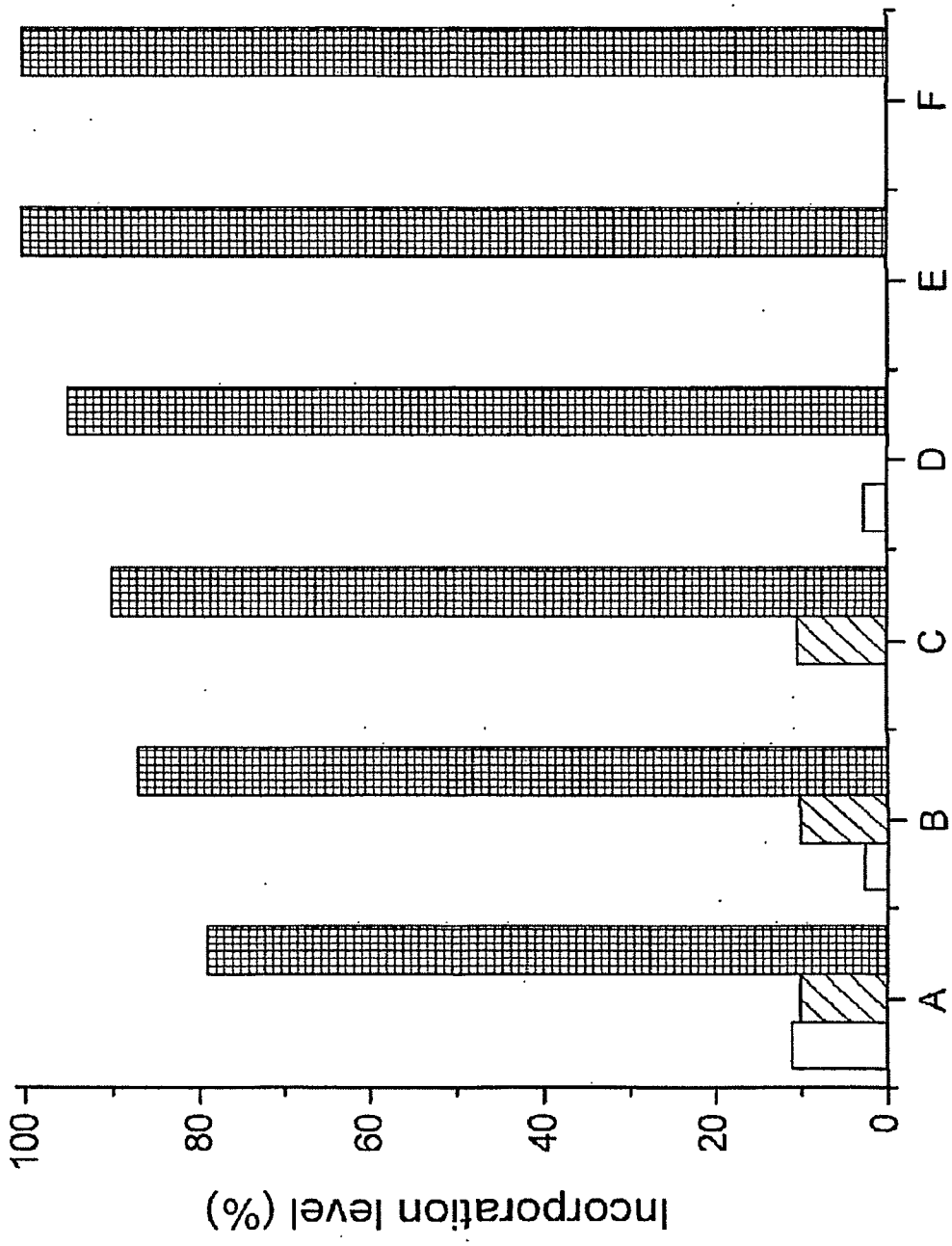
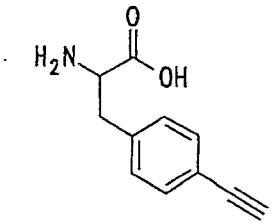


FIG. 12

Using yPheRS (T415G) variant and ytRNA^{Phe} amber suppressor, pEtF can be inserted into mDHFR in response to an amber codon in *E. coli* system.



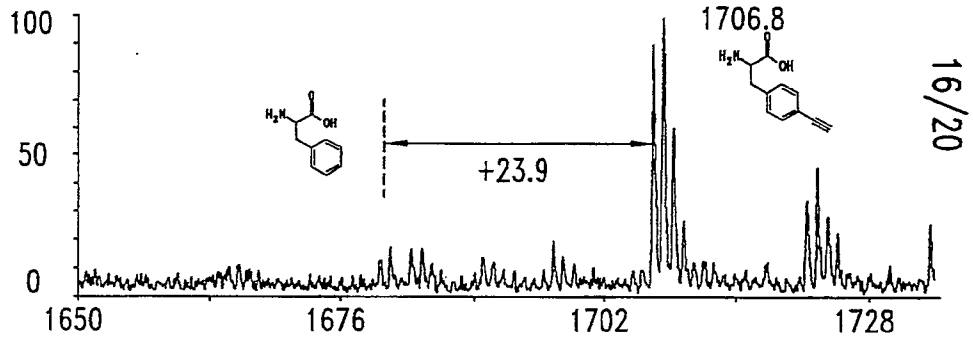
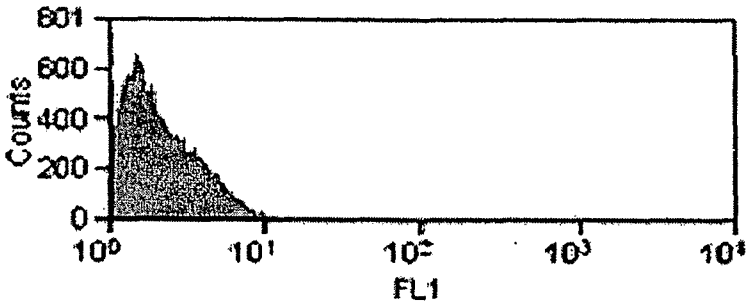


FIG. 13

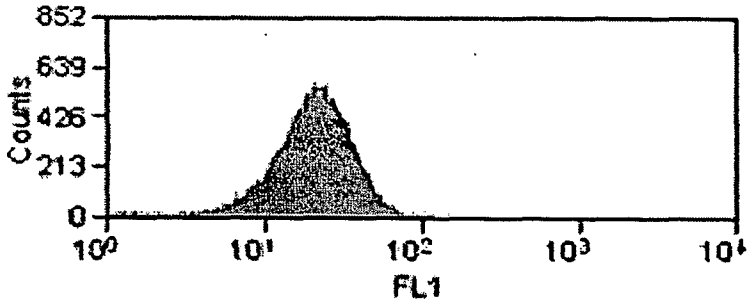
GFP amber suppression_081406

GFP_158 (L64, Am158)

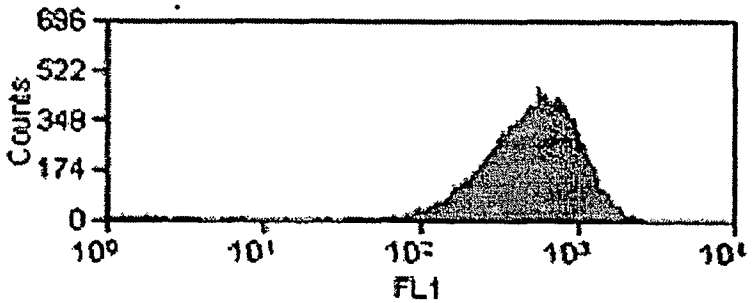
Before induction



0.01 mM Phe
0.01 mM Trp



0.05 mM Phe
0.05 mM Trp
3.0 mM 2Nal



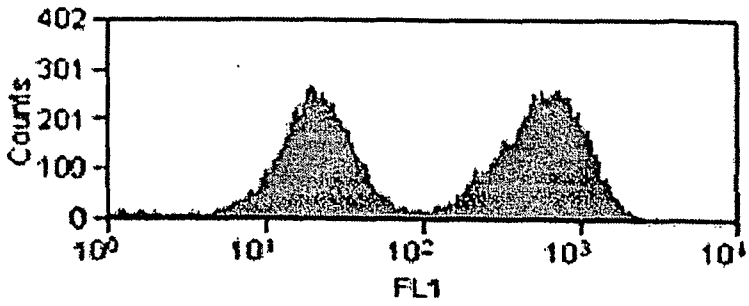
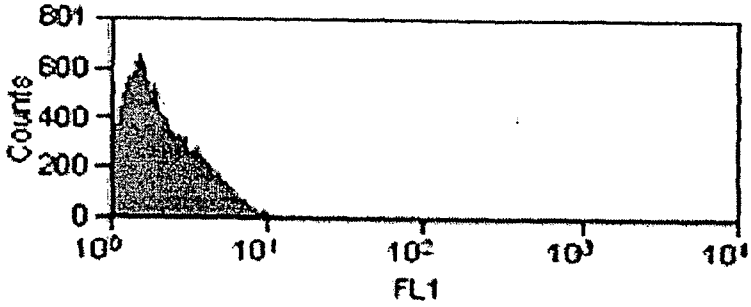


FIG. 14A

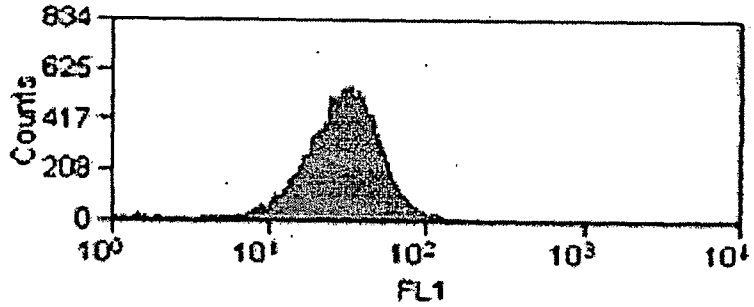
GFP amber suppression_081406

GFP_158 (L64, Am158)

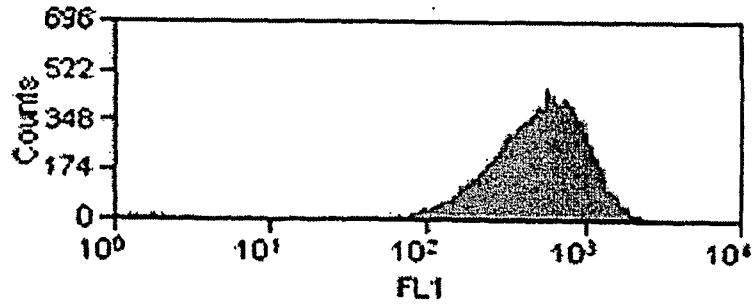
Before induction



0.05 mM Phe
0.05 mM Trp



0.05 mM Phe
0.05 mM Trp
3.0 mM 2Nal



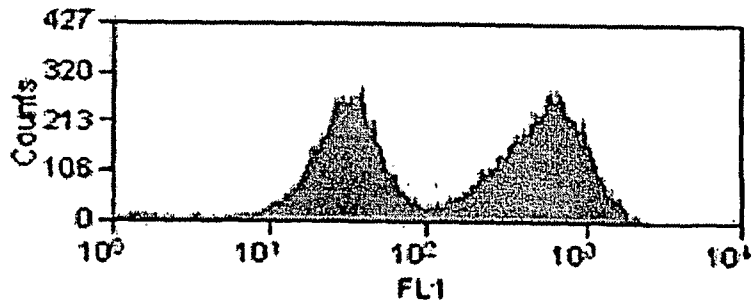


FIG. 14B

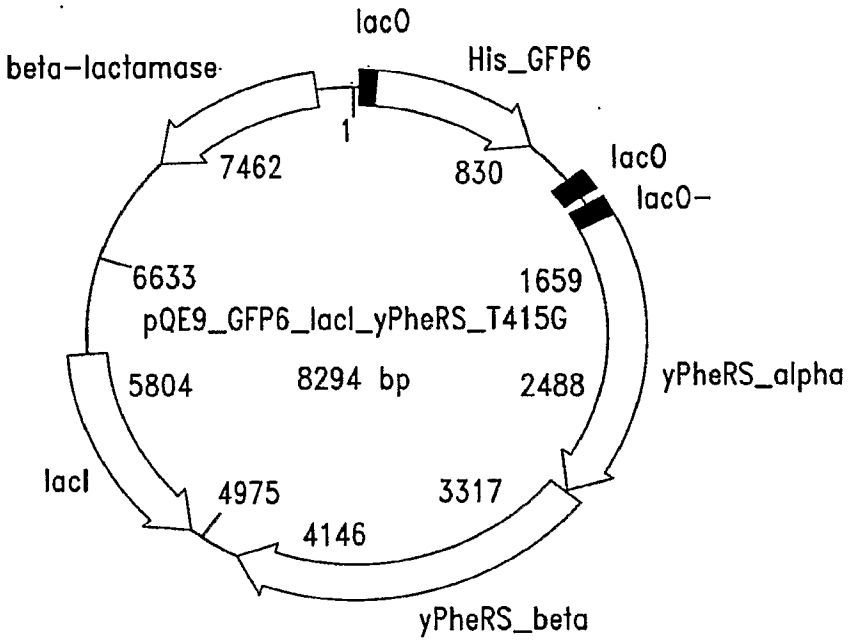

lacO
His_GFP6
beta-lactamase
1
7462
830
lacO
lacO-
6633
1659
pQE9_GFP6_lacI_yPheRS_T415G
5804
8294 bp
2488
yPheRS_alpha
lacI
4975
3317
4146
yPheRS_beta

FIG. 15

>yeast_PheRS_alpha_mutation_sites_are_marked_with_NNN

atgtctgact tccaattaga aattctaaag aaactagatg aattggatga gatcaagtcc
acactggcaa ctttccctca gcacggctct caagatgttc tttccgcttt gaactctttg
aaagcccaca acaagttaga gttttccaag gtcgacacgg ttacgtatga cttgaccaaa
gaaggtgctc aaattttgaa tgaaggttcg tacgaaatta aactagtcaa gctcatccaa
gagttgggtc aacttcaaat caaagatgtg atgtccaaac taggccctca agttggtaag
gtcggtcagg ctagagcttt caagaacggc tggatcgcca aaaacgcctc aaacgagctt
gaactctccg caaaattgca aaataccgat ttaaatgagc ttactgatga aacgcaatct
attctagcgc aaatcaagaa caactcgcat ctggatagca ttgacgccaa gattttgaac
gacttgaaga aaagaaagtt aattgctcaa ggtaaaatca cagatttcag tgtcaccaaa
gggccagagt tctcgaccga cctcaccaaa ttggaaaccg atcttacctc cgacatggtc
tccaccaatg catacaagga cttgaagttc aagccttaca atttcaattc tcaaggtgtg
caaatatctt caggtgctct tcacccctta aacaaagtca gagaggaatt tagacaaatt
ttcttttcca tgggattcac agagatgccc tcgaaccaat acgtcgagac aggtttctgg
aacttcgatg ccctttacgt cccacaacag catcctgctc gtgacctgca agacactttc
tacatcaagg acccactaac cgctgagttg cccgatgaca agacatacat ggacaatatc
aaagccgttc acgaacaggg gagattcggg tccatcggtt atcgttacaa ctggaagcca
gaagaatgtc aaaaattggt cttgagaact cactccacag ccatctctgc cagaatgctg
cacgatttgg ccaaagatcc aaagcccacc agattgtttt ctatcgaccg tgttttccgt
aacgaagcag ttgacgccac ccatttggcc gaattccacc aggtggaagg tgttcttgcc
gactacaaca ttactctggg tgacctgatc aagttcatgg aagagttttt cgaaagaatg
ggtgtcaccg gtttgagatt caagcctacc tacNNNcctt acNNNgagcc aNNNatggaa
atcttttctt ggcacgaagg tttgcaaaaa tgggtcgaaa tcggtaacNNNggtatgttc
agaccagaaa tgctcgagtc catgggtcta ccaaaggatc taagagtcct tggttggggg
ttatccttgg aaagacctac catgatcaaa tataaggttc aaaacatcag agaactgtta
ggtcataaag tctctttgga ctttatcgaa accaatcctg ctgctagatt ggacgaagac
ttgtacgaat aa

	Name	412	415	418	437
	T415G	AAT	GGC	TCA	TCT
1	2Nal	GGG	GGG	TGT	TTT
2	412_415	GGG	GGC		
3	415_418		GGC	TGT	
4	415_437		GGC		TTT
5	412_415_437	GGG	GGC		TTT
6	415_418_437		GGC	TGT	TTT
7	412_415_418	GGG	GGC	TGT	

FIG. 16