INNOVATION

A chemical toolkit for proteins — an expanded genetic code

Jianming Xie and Peter G. Schultz

Abstract | Recently, a method to encode unnatural amino acids with diverse physicochemical and biological properties genetically in bacteria, yeast and mammalian cells was developed. Over 30 unnatural amino acids have been co-translationally incorporated into proteins with high fidelity and efficiency using a unique codon and corresponding transfer-RNA:aminoacyl–tRNA-synthetase pair. This provides a powerful tool for exploring protein structure and function *in vitro* and *in vivo*, and for generating proteins with new or enhanced properties.

Although the genetic codes of all known organisms specify the same 20 amino acids (with the rare exceptions of selenocysteine[1] and pyrrolysine[2]), it is clear that numerous proteins require many cofactors and post-translational modifications to carry out their natural functions. Therefore, although a 20-amino-acid code might be sufficient for life, it might not be optimal. Consequently, the development of a method that allows us to encode extra amino acids genetically might facilitate the evolution of proteins, or even entire organisms, with new or enhanced properties. Moreover, the ability to incorporate amino acids with defined steric and electronic properties at unique sites in proteins will provide powerful new tools for exploring protein structure and function *in vitro* and *in vivo*.

> “...the ability to incorporate amino acids with defined steric and electronic properties at unique sites in proteins will provide powerful new tools for exploring protein structure and function *in vitro* and *in vivo*.”

Here, we describe an approach that makes it possible, for the first time, to add new amino acids to the genetic codes of both prokaryotic and eukaryotic organisms. Over 30 unnatural amino acids — including those containing spectroscopic probes, post-translational modifications, metal chelators, photoaffinity labels and other chemical moieties — have been selectively incorporated into proteins with high fidelity and efficiency in response to unique three and four base codons.

Methodology

General considerations. The incorporation of an unnatural amino acid at a defined site in a protein, directly in a living organism, requires a unique transfer-RNA:codon pair, a corresponding aminoacyl–tRNA synthetase and significant intracellular levels of the unnatural amino acid[3]. To ensure that the unnatural amino acid is incorporated uniquely at the site specified by its codon, the tRNA must be constructed such that it is not recognized by the endogenous aminoacyl–tRNA synthetases of the host, but functions efficiently in translation (an orthogonal tRNA). Moreover, this tRNA must deliver the novel amino acid in response to a unique codon that does not encode any of the common 20 amino acids. Another requirement for high fidelity is that the cognate aminoacyl–tRNA synthetase (an orthogonal synthetase) aminoacylates the orthogonal tRNA, but does not aminoacylate any of the endogenous tRNAs. Furthermore, this synthetase must aminoacylate the tRNA with only the desired unnatural amino acid and not with the endogenous amino acids. Similarly, the unnatural amino acid cannot be a substrate for the endogenous synthetases if it is to be incorporated uniquely in response to its cognate codon. Last, the unnatural amino acid must be efficiently transported into the cytoplasm when it is added to the growth medium or biosynthesized by the host, and it must be stable in the presence of endogenous metabolic enzymes.

Several biochemical methods have previously been developed to insert unnatural amino acids into proteins. However, they require either *in vitro* protein synthesis or the stoichiometric use of chemically aminoacylated tRNAs[4–6] (which results in low protein yields), or they result in the substitution of an unnatural amino acid (which typically must be a close structural analogue of a common amino acid) throughout the proteome or the partial incorporation of the unnatural amino acid in competition with endogenous amino acids[7–10]. The challenge is to develop a general method that makes it possible to incorporate a wide range of unnatural amino acids at any genetically specified site in the proteome with high translational fidelity and efficiency.

Encoding unnatural amino acids in prokaryotes. Initially, the amber nonsense codon (UAG) was used to specify a novel amino acid in *Escherichia coli* because it is the least used stop codon and the presence of natural amber suppressors in some *E. coli* strains does not significantly affect cell-growth rates[11,12]. To obtain orthogonal-tRNA:aminoacyl–tRNA-synthetase pairs that uniquely encode extra amino acids in bacteria, orthologues were taken from archaea. The first such pair was derived from the tyrosyl–tRNA:tyrosyl–tRNA-synthetase pair from *Methanococcus jannaschii* (*Mj*tRNA^{Tyr}:*Mj*TyrRS)[13]. *Mj*tRNA^{Tyr} has distinct synthetase recognition elements compared to bacterial tRNAs[14] (FIG. 1a), and the cognate synthetase *Mj*TyrRS can be expressed efficiently in *E. coli*. Also, *Mj*TyrRS has a minimalist anticodon-loop-binding domain[15], which made it possible to alter the anticodon loop of *Mj*tRNA^{Tyr} to CUA with a minimal reduction in its affinity for the synthetase. Last, *Mj*TyrRS does not have an editing mechanism that could deacylate the unnatural amino acid[16].

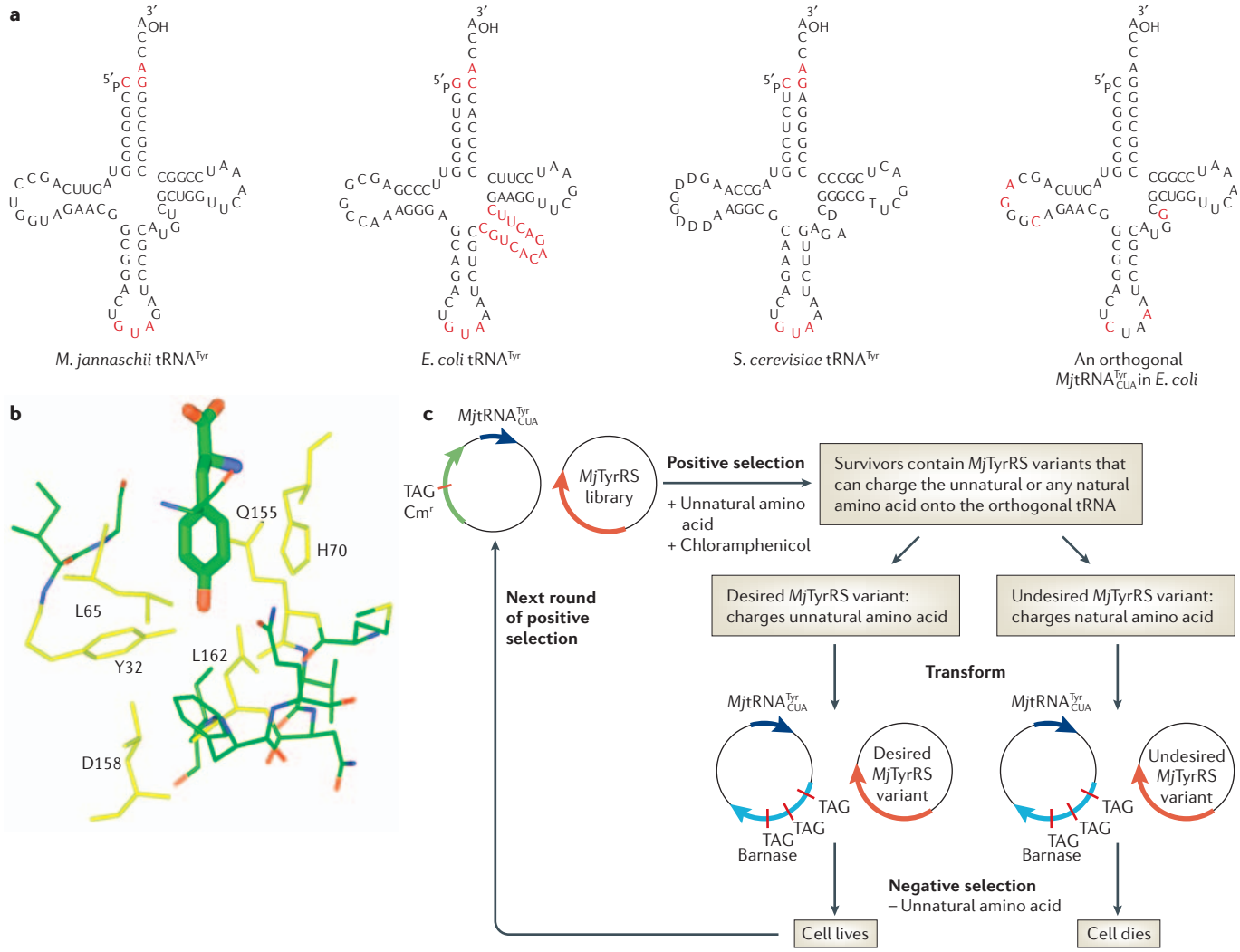

Figure 1 | **Encoding unnatural amino acids in prokaryotes.** The development of an orthogonal amber suppressor *Methanococcus jannaschii* tyrosyl–transfer-RNA (*Mj*tRNA$^{Tyr}_{CUA}$) in *Escherichia coli* and the modification of the amino-acid specificity of its cognate *M. jannaschii* tyrosyl–tRNA synthetase (*Mj*TyrRS). **a** | The tRNA^{Tyr} molecules from *M. jannaschii*, *E. coli* and *Saccharomyces cerevisiae* (with the key identity elements that are recognized by the cognate synthetases highlighted in red), and the orthogonal amber suppressor *Mj*tRNA$^{Tyr}_{CUA}$ in *E. coli* (with the modified nucleotides highlighted in red). The D nucleotide is dihydrouridine. **b** | A library of *Mj*TyrRS mutants was generated by randomly mutating 6 residues (shown in yellow) in the Tyr-binding site to all 20 amino acids. Tyr is shown in its binding site using a thicker stick representation. **c** | A general positive and negative selection scheme for the development of synthetase variants that are specific for an unnatural amino acid in *E. coli*. Following the generation of a large library (~10^9 mutants) of, in this case, *Mj*TyrRS active-site mutants, positive and negative selections were carried out. The positive selection was based on resistance to chloramphenicol, which was conferred in the presence of *Mj*TyrRS and the unnatural amino acid (or any natural amino acid that the *Mj*TyrRS could charge onto the orthogonal tRNA) by the suppression of an amber mutation (TAG) at a permissive site in the chloramphenicol acetyltransferase gene (labelled Cm^r). The negative selection used the toxic barnase gene with amber mutations at permissive sites and was carried out in the absence of the unnatural amino acid. Only *Mj*TyrRS variants that could acylate the orthogonal tRNA$^{Tyr}_{CUA}$ with the unnatural amino acid and not with the endogenous amino acids could survive both selections.

Unfortunately, the amber suppressor *Mj*tRNA$^{Tyr}_{CUA}$ was found to be aminoacylated to some degree by the endogenous *E. coli* synthetases. To improve the orthogonality of *Mj*tRNA$^{Tyr}_{CUA}$, 11 nucleotides that do not interact directly with *Mj*TyrRS were randomly mutated to generate a suppressor-tRNA library[17]. *Mj*tRNA$^{Tyr}_{CUA}$ variants that were substrates for endogenous *E. coli* aminoacyl–tRNA synthetases were removed from the library using a round of negative selection that was based on the suppression of amber nonsense mutations in the toxic ribonuclease barnase gene. To select functional orthogonal-tRNA:aminoacyl–tRNA-synthetase pairs, a round of positive selection was used that was based on the capability of *Mj*tRNA$^{Tyr}_{CUA}$ variants to confer ampicillin resistance by suppressing an amber mutation at a permissive site in the β-lactamase gene when the cognate *Mj*TyrRS was present. Alternating negative and positive rounds of selection led to the identification of a mutant *Mj*tRNA$^{Tyr}_{CUA}$ (FIG. 1a) that is not a substrate for endogenous *E. coli* synthetases, that can be aminoacylated by *Mj*TyrRS and that functions efficiently in translation.

Next, it was necessary to alter the substrate specificity of the orthogonal *Mj*TyrRS so that it recognized the unnatural amino acid and not Tyr or any other endogenous amino acids. A general

approach[18] was developed that involved generating a large library (~10^9 mutants) of synthetase active-site mutants[19,20] (FIG. 1b) followed by a combination of positive and negative selections to identify a synthetase with the desired specificity (FIG. 1c). The positive selection was based on resistance to chloramphenicol, which, in the presence of the unnatural amino acid and *Mj*TyrRS, was conferred by the suppression of an amber mutation at a permissive site in the chloramphenicol acetyltransferase gene. The negative selection used the toxic barnase gene with amber mutations at permissive sites and was carried out in the absence of the unnatural amino acid. Only *Mj*TyrRS variants that could acylate the orthogonal *Mj* $tRNA^{Tyr}_{CUA}$ with the unnatural amino acid and not with the endogenous amino acids could survive both selections.

This selection scheme and more recent variants have been used to develop *Mj*TyrRS mutants that are capable of selectively inserting over 30 unnatural amino acids into proteins in *E. coli* in response to the amber codon[3]. Typically, 5–10 mg l^{-1} of an unnatural-amino-acid-containing protein can be obtained from minimal media with translational fidelities of over 99%. Recently the system was optimized, which has made it possible to produce approximately 500 mg l^{-1} of proteins that contain unnatural amino acids in bacteria using high-density fermentation[21]. Meanwhile, further orthogonal suppressor-tRNA:aminoacyl–tRNA-synthetase pairs have been generated[22–25], which increase the structural diversity and number of unnatural amino acids that can be incorporated into proteins using this method. Moreover, we have shown that it is possible to add an engineered pathway for the biosynthesis of an unnatural amino acid (*p*-aminophenylalanine) into *E. coli* to generate an autonomous '21-amino-acid' bacterium[26].

Encoding unnatural amino acids in eukaryotes. A similar strategy has been used to generate nonsense-suppressor-tRNA:aminoacyl–tRNA-synthetase pairs in *Saccharomyces cerevisiae* from *E. coli* orthologues[27,28]. In this case, a positive and negative selection scheme was developed that used the transcriptional activator Gal4 with two amber mutations at permissive sites. Suppression of these amber codons led to the production of full length Gal4, which, in turn, drove the transcription of positive or negative selection reporters. This straightforward selection scheme together with orthogonal-tRNA:aminoacyl–tRNA-synthetase pairs — including *E. coli* $tRNA^{Tyr}_{CUA}$:TyrRS and $tRNA^{Leu}_{CUA}$:leucyl–tRNA-synthetase pairs — has allowed us to incorporate over 15 unnatural amino acids into proteins in *S. cerevisiae*. Expression levels of up to 75 mg l^{-1} have been obtained with greater than 98% fidelities.

More recently, the mutant aminoacyl–tRNA synthetases that were evolved in *S. cerevisiae* to accept unnatural amino acids have been used together with a *Bacillus stearothermophilus* amber suppressor $tRNA^{Tyr}_{CUA}$ to selectively insert various unnatural amino acids into proteins in mammalian cells in response to nonsense codons (W. Liu and P.G.S., unpublished results). Also, Yokoyama and co-workers screened a collection of designed active-site variants of *E. coli* TyrRS in a wheat-germ translation system and discovered a mutant synthetase that uses 3-iodotyrosine more effectively than Tyr[29]. This mutant synthetase was later used with *B. stearothermophilus* $tRNA^{Tyr}_{CUA}$ to incorporate 3-iodotyrosine into proteins in mammalian cells[30]. Similarly, we used a mutant orthogonal *Bacillus subtilis* $tRNA^{Trp}_{UCA}$:tryptophanyl–tRNA-synthetase pair to introduce the redox-active crosslinking agent 5-hydroxytryptophan selectively into proteins in response to the UGA opal codon[31]. However, it remains a challenge to produce large quantities of proteins containing unnatural amino acids in mammalian cells.

> "...we have shown that it is possible to add an engineered pathway for the biosynthesis of an unnatural amino acid ... into *E. coli* to generate an autonomous '21-amino-acid' bacterium."

Further codons for unnatural amino acids. To incorporate two or more distinct unnatural amino acids into a single protein simultaneously, further unique codons (other than the amber and opal nonsense codons) are needed. It should be possible to use quadruplet codons and cognate suppressor tRNAs with expanded anticodon loops to specify unnatural amino acids[32,33]. Recently, we showed that an orthogonal $tRNA^{Lys}_{UCCU}$:lysyl–tRNA-synthetase pair from *Pyrococcus horikoshii* could be used to incorporate homoglutamine selectively into proteins in *E. coli* in response to the quadruplet codon AGGA. The combination of amber and AGGA suppressions allowed the simultaneous incorporation of two unnatural amino acids at distinct sites in a single protein[23]. An alternative approach to generate further codons involves eliminating degenerate nonsense codons and codon–tRNA pairs from the *E. coli* genome (avoiding competition with stop or coding codons). We are currently evaluating methods for the efficient construction of a partially 'codon-deleted' *E. coli* genome. A remaining challenge is to define the role of context effects in suppression efficiency[34] (the efficiency with which nonsense codons can be suppressed can be affected by the characteristics of adjacent codons) and to identify genomic mutations (for example, in the ribosome, tRNAs and elongation factors) that improve suppression efficiency.

An expanded genetic code

The above methodology has been successfully used to add a large number of diverse unnatural amino acids to the genetic codes of *E. coli*, *S. cerevisiae* and mammalian cells. Many of these unnatural amino acids have novel properties that are useful for various biochemical and cellular studies of protein structure and function (FIG. 2).

Chemically reactive groups. A potential general strategy for selective protein modification involves the site-specific incorporation of unnatural amino acids with novel reactivity into proteins, which can subsequently be derivatized with high efficiency and selectivity. Indeed, several unnatural amino acids with reactive groups, including ketone, azide, acetylene and thioester groups, have been genetically encoded in *E. coli* and *S. cerevisiae*[27,35–39] (1–8 in FIG. 2). These chemistries have been used to modify proteins selectively under mild conditions with a number of fluorophore[35,36,40] tags and other exogenous reagents. In one example, a series of fluorescent dyes were selectively introduced at a unique site in an *m*-acetylphenylalanine mutant of the membrane protein LamB in *E. coli*[36]. In a second example, a mutant human growth hormone was site-specifically modified with polyethylene glycol (PEG) with a high yield to create a protein that retained wild-type activity but that had a considerably improved half-life in serum (H. Cho and T. Daniel, unpublished results). This work is currently being extended to other therapeutic proteins, as well as to the generation of homodimeric and heterodimeric

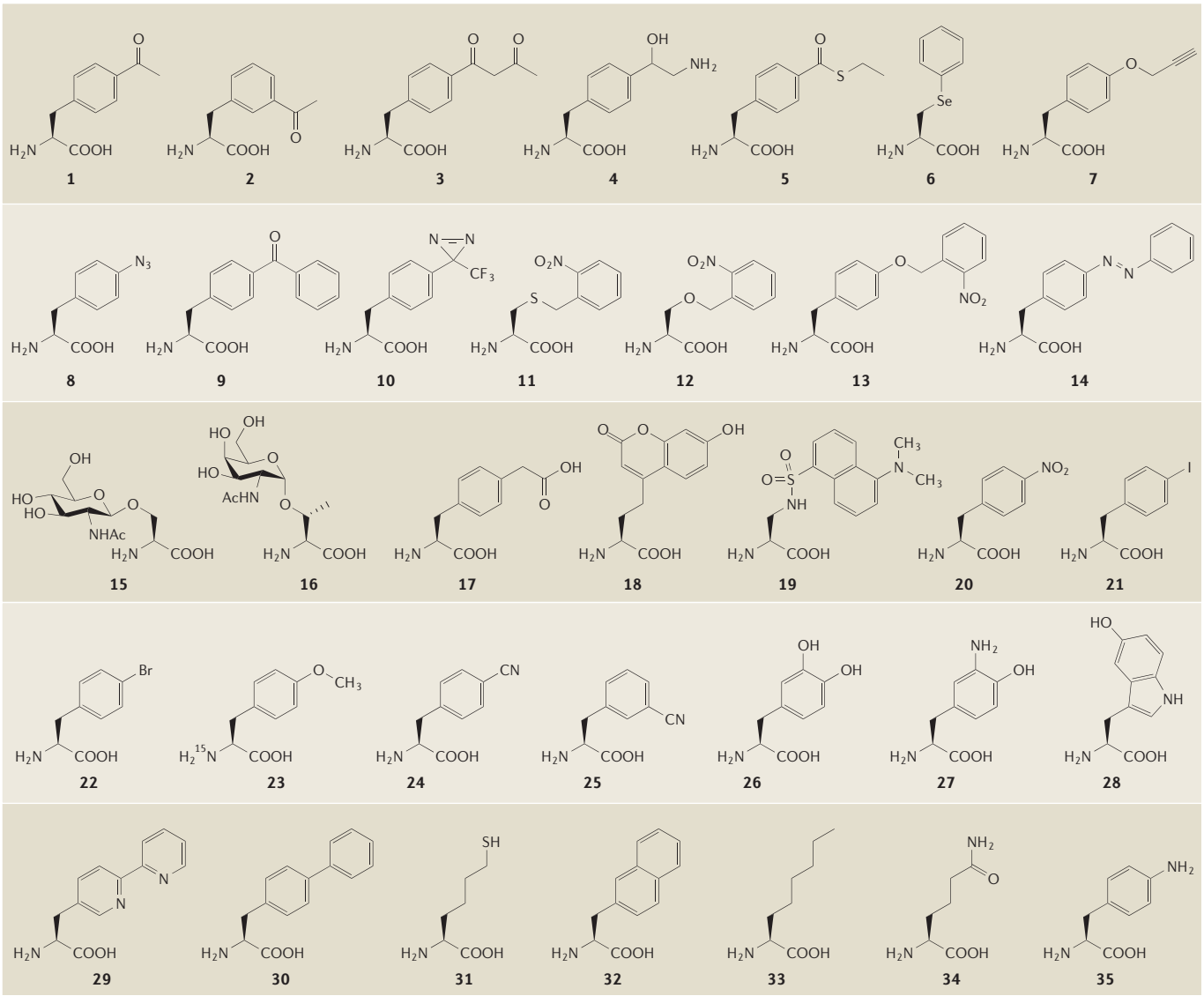

Figure 2 | **Unnatural amino acids that have been added to the genetic codes of prokaryotes and eukaryotes. 1–8** | Unnatural amino acids with uniquely reactive groups that can be used to modify proteins selectively (the reactive groups are: **1** and **2**, ketones; **3**, β-diketone; **4**, 1,2-hydroxyamine; **5**, thioester; **6**, phenylselenide; **7**, acetylene; and **8**, azide). **8–14** | Unnatural amino acids with photoreactive side chains (**8–10**, photocrosslinkers for mapping protein–protein and protein–nucleic-acid interactions; **11–13**, caged amino acids for the photoactivation of protein function; and **14**, a photochromic amino acid that can be used to reversibly regulate protein function). **15–17** | Unnatural amino acids with post-translational modifications (**15** and **16**, glycosylated residues; and **17**, a phosphatase-stable analogue of phosphotyrosine). **18–25** | Biophysical probes (**18** and **19**, fluorescent probes; **20**, a fluorescence quencher; **21** and **22**, amino acids containing heavy atoms (for protein phasing); **23**, an ^{15}N NMR probe; and **24** and **25**, infrared probes). **26–28** | Redox-active amino acids (**26** and **27**, probes of electron-transfer processes or radical traps; and **28**, an electrochemical crosslinker). **29** | A metal chelator. **30–35** | Other representative unnatural amino acids (**30** and **32**, aromatic hydrophobic amino acids that could be used to modulate hydrophobic packing interactions; **31**, a long-chain Cys analogue that can be used to form extended disulphide bonds; **33**, an amino acid with a long aliphatic side chain; **34**, L-homoglutamine; and **35**, *p*-aminophenylalanine).

PEG-linked Fab fragments, and it should allow the synthesis of chemically modified proteins with unprecedented control over their structure and homogeneity. It should also be useful for the ribosomal synthesis of cyclic peptides and proteins. For example, if a ketone is incorporated at the C terminus of a protein, it can subsequently react with the N-terminal amine group to cyclize the protein.

Photoreactive groups. The introduction of photocrosslinkers at unique sites in proteins *in vivo* should provide a powerful tool to map biomolecular interactions. To this end, aminoacyl–tRNA-synthetase mutants have been evolved that can selectively incorporate three unnatural amino acids with photoreactive side chains into proteins[27,37,41–43] — *p*-azidophenylalanine, *p*-benzoylphenylalanine and *p*-(3-trifluoromethyl-3*H*-diazirin-3-yl)-phenylalanine (8–10 respectively in FIG. 2). Among these, *p*-benzoylphenylalanine is particularly useful because the benzophenone moiety absorbs light at relatively long wavelengths (~360 nm) and turns into a long-lived triplet state that inserts efficiently into C–H bonds[44]. Indeed, a mutant homodimeric glutathione *S*-transferase with *p*-benzoylphenylalanine substituted site-specifically at the

dimer interface could be crosslinked in the cytoplasm of *E. coli* cells with a greater than 50% yield[42]. This amino acid was recently used as a substitute for residues in the central pore site of ClpB, a ring-forming AAA+ (ATPases associated with various cellular activities) protein that rescues proteins from aggregated states. The resulting mutants were shown to photocrosslink to the peptide substrate, which provided direct physical evidence for close contacts between the ClpB pore site and substrates[45,46]. Yokoyama and co-workers also showed that *p*-benzoylphenylalanine can be selectively incorporated into human GRB2 (growth-factor-receptor-bound protein-2) in Chinese hamster ovary cells and crosslinked to the epidermal-growth-factor receptor[47]. These amino acids should be useful as probes of protein interactions, protein structure and protein dynamics *in vitro* and *in vivo*, and they should also be helpful in identifying receptors for orphan ligands.

Photocaged Cys, Ser and Tyr amino acids (11–13 respectively in FIG. 2) have also been site-specifically introduced into proteins using this methodology[28,48]. The side-chain hydroxy or thiol groups of these amino acids are blocked by substituted nitrobenzyl groups that can be cleaved on irradiation with 365-nm light *in vitro* or *in vivo*. In one example, the active-site Cys of the pro-apoptotic cysteine protease caspase-3 was substituted with nitrobenzyl Cys to create an inactive protein that could be photoactivated with greater than 70% efficiency. In a related experiment, the photochromic[49] amino acid *p*-azophenyl-phenylalanine (14 in FIG. 2) was site-specifically introduced into the cyclic-AMP-binding site of the *E. coli* transcription factor catabolite activator protein (CAP)[50]. Irradiation of this amino acid with 334-nm light predominantly converts the *trans* form of the amino acid to the *cis* form. Because the two isomers differ significantly in structure and dipole, the *cis* and *trans* mutants have different affinities for cAMP and, as a result, the affinity of this mutant CAP for its promoter can be photoregulated. Similarly, it should be possible to photomodulate the activity of other enzymes (for example, kinases and phosphatases), receptors and transcription factors.

Post-translational modifications. The difficulties that are associated with the generation of selectively glycosylated proteins have hindered our understanding of the biological roles of glycosylation and have also made the production of therapeutically useful glycoproteins challenging[51]. To begin to provide a general approach for the synthesis of structurally-defined glycoproteins, mutant synthetases have been evolved that site-specifically incorporate β-*N*-acetylglucosamine-*O*-serine (β-GlcNAc-Ser; 15 in FIG. 2) into proteins in *E. coli*[52]. It was shown that a β-GlcNAc-Ser-containing mutant myoglobin that was generated by this method could be further modified by galactosyltransferases to produce more complex saccharides. A similar approach was used to introduce α-*N*-acetylgalactosamine-*O*-threonine (16 in FIG. 2) selectively into proteins[53], and is currently being extended to a number of other *O*- and *N*-linked sugars.

Reversible protein phosphorylation, principally on Ser, Thr or Tyr residues, is crucial in the regulation of signal-transduction pathways. Consequently, the generation of selectively phosphorylated proteins or stable analogues of phosphoproteins would be useful[54]. As a first step, we have selectively incorporated a non-phosphorus-containing analogue of phosphotyrosine — *p*-carboxymethyl-L-phenylalanine (17 in FIG. 2) — into proteins in *E. coli*. When this unnatural amino acid is used to replace Tyr701 in human STAT1 (signal transducer and activator of transcription-1), the resulting protein dimerizes and binds to the same DNA sequence as Tyr701-phosphorylated STAT1 (REF. 55). Because this amino acid has better cellular membrane permeability than phosphotyrosine and is resistant to protein tyrosine phosphatases, it should be useful in the generation of other stable phosphoprotein analogues or peptide-based inhibitors for SH2 (Src-homology-2) domains.

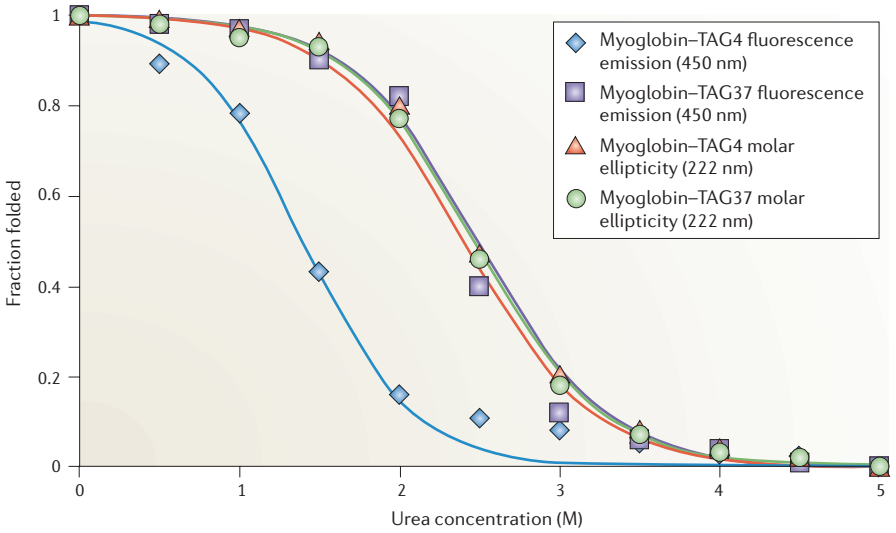

Figure 3 | **The site-specific incorporation of a coumarin-derived fluorescent amino acid into myoglobin as a probe of protein conformational changes.** An orthogonal *Methanococcus jannaschii* tyrosyl–transfer-RNA ($MjtRNA^{Tyr}_{CUA}$):tyrosyl–tRNA-synthetase pair was identified that specifically incorporates a coumarin-derived amino acid (18 in FIG. 2) into proteins in response to the amber TAG codon. Because coumarin fluorescence is sensitive to solvent polarity, its fluorescence intensity should correlate with any local unfolding of the protein that occurs in close proximity to the mutated position. To illustrate this, the coumarin-derived amino acid was incorporated at position 4 or 37 of sperm whale myoglobin (myoglobin–TAG4 or myoglobin–TAG37, respectively). The resulting mutants were unfolded by urea, and this was monitored by measuring the fluorescence intensity of the incorporated coumarinyl amino acid and by circular dichroism (CD)[56]. In the presence of 5 M urea (fully unfolded state), both of the myoglobin mutants showed a 30% increase in their fluorescence signal compared to that at 0 M urea (fully folded state), which indicates that the fluorescence intensity of coumarin correlates with protein conformational changes. The 'fraction folded' was calculated by normalizing the fluorescence signal (or the molar ellipticity, which is an experimental parameter for CD) at 5 M urea to the fully unfolded state. On moving from 0 M to 2 M urea, the fluorescence intensity of myoglobin–TAG4 increased by 25% (and remained roughly at this level from 2 M to 5 M urea), which indicates that this region of the protein is disordered. By contrast, myoglobin–TAG37 showed little change in its fluorescence intensity at 2 M urea, but underwent a similar fluorescence increase (~25%) at 3 M urea. This is consistent with NMR data[68], which indicate that helices A and B of myoglobin are largely disordered when the urea concentration is higher than 2.2 M (helix A contains residue 4), whereas helices C, D and F unfold when the urea concentration is higher than 3.0 M (helix C contains residue 37). It therefore seems that the coumarinyl amino acid is a site-specific probe of protein conformational changes. The CD measurements produced virtually identical unfolding curves for myoglobin–TAG4 and myoglobin–TAG37, which is consistent with the efficacy of CD in reporting global conformational changes that are averaged over the entire structure.

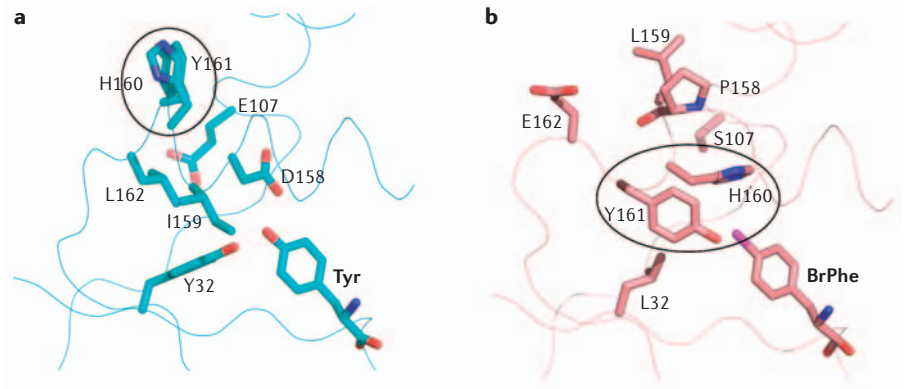

Figure 4 | **The structures of the wild-type and a mutant *Methanococcus jannaschii* tyrosyl–tRNA synthetase bound to their cognate amino acids. a** | The active site of wild-type *Methanococcus jannaschii* tyrosyl–transfer-RNA synthetase (*Mj*TyrRS) bound to Tyr. **b** | The active site of a mutant *Mj*TyrRS that binds to *p*-bromophenylalanine (labelled BrPhe in the figure). The active site of the mutant contains the mutations Y32L, E107S, D158P, I159L and L162E. The active-site D158P and Y32L mutations remove two hydrogen bonds to the hydroxyl group of the Tyr side chain, which disfavours the binding of the natural substrate. The D158P mutation results in the termination of helix α8 and produces significant translational and rotational movements of several active-site residues. These effects, in conjunction with the effects of the Y32L mutation, lead to an expanded hydrophobic active-site cavity that favours the binding of *p*-bromophenylalanine. Black frames highlight the different positioning of H160 and Y161 in these structures.

Biophysical probes. Unnatural amino acids that can function as probes of protein structure and function *in vitro* and *in vivo* have also been genetically encoded in *E. coli* and *S. cerevisiae*. Fluorescent amino acids (18 and 19 in FIG. 2) with 7-hydroxycoumarin and dansyl side chains have been selectively incorporated into proteins[56,57]. These unnatural amino acids provide small fluorescent probes for the direct visualization of protein conformational changes, localization and intermolecular interactions. They also offer an advantage over green fluorescent protein (GFP) and its derivatives by virtue of their smaller size and the fact that they can be introduced anywhere in a protein rather than just at the N or C terminus. Indeed, the coumarin derivative (18 in FIG. 2) has been used as a probe of local unfolding in myoglobin[56] (FIG. 3). It should be possible to incorporate other longer wavelength fluorophores as well. The unnatural amino acid *p*-nitrophenylalanine (20 in FIG. 2) has also been selectively incorporated into proteins in *E. coli* and can be used as a distance probe by quenching the intrinsic fluorescence of tryptophan[58]. The heavy-atom-containing amino acid *p*-iodo-L-phenylalanine (21 in FIG. 2) was genetically encoded in both *E. coli* and *S. cerevisiae*, and was used for single-wavelength anomalous diffraction phase determination in X-ray crystallography[59] (22 in FIG. 2 is another heavy-atom-containing amino acid that can also be used for protein phasing). ^{15}N-labelled *O*-methyltyrosine (23 in FIG. 2) has also been selectively incorporated into proteins as a site-selective NMR probe[60]. Last, *p*-cyanophenylalanine and *m*-cyanophenylalanine (24 and 25 respectively in FIG. 2) have been successfully incorporated into proteins and, because the nitrile group has a distinct vibrational frequency relative to the other functional groups that are present in proteins, they should be useful infrared probes for the local environment and protein dynamics[61].

Redox-active unnatural amino acids. Electron-transfer phenomena are involved in a large number of biological activities that range from enzyme catalysis to the primary charge-separation processes in photosynthesis. Two redox-active amino acids — dihydroxyphenylalanine and 3-amino-L-tyrosine (26 and 27 respectively in FIG. 2) — have been selectively incorporated into proteins[62]. They can undergo two-electron oxidation to form the corresponding quinone, and can be used to probe and manipulate electron-transfer processes in proteins. Furthermore, 3-amino-L-tyrosine can also function as a radical trap owing to the stability of its oxidized semiquinone form. These amino acids require reducing agents to be supplied in the growth media. Unnatural amino acids can also be used to perturb the electronic properties of a protein. For example, the introduction of several unnatural amino acids at Tyr66 in GFP led to changes in the absorbance and fluorescence spectra for this protein and to altered quantum yields[63].

Metal-chelating amino acids. The creation of proteins with engineered metal-binding sites remains a challenge owing to a difficulty in predicting and controlling the primary and secondary amino-acid shells that surround the metal ion. A potential solution is to expand the repertoire of protein building blocks to include multidentate metal-chelating amino acids, in which several ligating atoms are pre-orientated in the correct geometry. As a first step in this direction, we attempted to incorporate bipyridyl-alanine (29 in FIG. 2) selectively into proteins in *E. coli* in response to the amber nonsense codon. In this case, we could not evolve a synthetase directly, but rather first evolved a synthetase that was specific for a biphenyl analogue (30 in FIG. 2). The production of a second generation library from this synthetase, in conjunction with subsequent genetic selections, resulted in a synthetase that is specific for bipyridyl-alanine[55]. The incorporation of a number of other bidentate and tridentate metal-chelating amino acids is currently work in progress. These metal-chelating amino acids should be useful when incorporated into proteins in processes such as catalysis, metal-ion-dependent protein dimerization, the formation of fluorescent metal–ligand complexes and the binding of radionuclides.

Other representative unnatural amino acids. In another series of experiments, a long-chain analogue of Cys (31 in FIG. 2) has been selectively incorporated into proteins in *S. cerevisiae*. This approach is currently being extended to a number of Cys analogues with different side-chain lengths, which might be useful in stabilizing specific protein conformations, studying the dynamics of protein folding or designing thermostable proteins. Other representative unnatural amino acids that have been added to the genetic code of *E. coli* or *S. cerevisiae* include β-(2-naphthyl)alanine[64], α-aminocaprylic acid[28], L-homoglutamine and *p*-aminophenylalanine[26] (32–35 respectively in FIG. 2). Clearly, the large number of structurally diverse amino acids that have been genetically encoded so far indicates that the translational machinery is rather tolerant of side-chain structure. Indeed, X-ray crystallographic studies have shown that

there is a great deal of structural plasticity in the conformations of the amino-acid side chains and the protein backbone in the amino-acid-binding sites of these aminoacyl–tRNA synthetases[65,66] (FIG. 4). It should therefore be possible to add other novel amino acids to the genetic code including spin labels, electron-transfer mediators, near-infrared probes and long-chain alkanes, as well as building blocks with altered backbones such as α-hydroxy acids and *N*-alkyl amino acids.

Conclusions

The approach described above has proven remarkably effective in allowing us to add a large number of structurally diverse amino acids to the genetic codes of both prokaryotic and eukaryotic organisms. The ability to encode unnatural amino acids genetically should provide powerful probes of protein structure and function *in vitro* and *in vivo*. It might also allow the design or evolution of proteins with novel properties. Possible examples include the rational design of glycosylated or 'PEGylated' therapeutic proteins with improved pharmacological properties, fluorescent proteins that function as sensors of small molecules and protein–protein interactions in cells, and proteins with activities that can be photoregulated *in vivo*. We might also be able to select for peptides and proteins that have enhanced function using libraries of unnatural-amino-acid mutants. For example, we recently showed that it is possible to incorporate unnatural amino acids into phage-displayed peptides[67], and a peptide with an enhanced affinity for streptavidin was isolated and found to contain an unnatural amino acid. It should also be possible to incorporate non-amino-acid building blocks into proteins or perhaps to create biopolymers with entirely unnatural backbones, as well as to generate multicellular organisms that contain unnatural amino acids. Last, the capability to add novel amino acids to the genetic code of organisms should allow us to test experimentally whether organisms with genetic codes that contain more than twenty amino-acid building blocks have an evolutionary advantage.

Jianming Xie and Peter G. Schultz are at the Department of Chemistry and Skaggs Institute for Chemical Biology, the Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
Correspondence to P.G.S.
e-mail: schultz@scripps.edu

doi:10.1038/nrm2005
Published online 23 August 2006

1. Bock, A. *et al.* Selenocysteine: the 21st amino acid. *Mol. Microbiol.* **5**, 515–520 (1991).
2. Srinivasan, G., James, C. M. & Krzycki, J. A. Pyrrolysine encoded by UAG in archaea: charging of a UAG-decoding specialized tRNA. *Science* **296**, 1459–1462 (2002).
3. Wang, L. & Schultz, P. G. Expanding the genetic code. *Angew. Chem. Int. Edn Engl.* **44**, 34–66 (2004).
4. Cornish, V. W., Mendel, D. & Schultz, P. G. Probing protein structure and function with an expanded genetic code. *Angew. Chem. Int. Edn Engl.* **34**, 621–633 (1995).
5. Bain, J. D., Glabe, C. G., Dix, T. A., Chamberlin, A. R. & Diala, E. S. Biosynthetic site-specific incorporation of a non-natural amino acid into a polypeptide. *J. Am. Chem. Soc.* **111**, 8013–8014 (1989).
6. Beene, D. L., Dougherty, D. A. & Lester, H. A. Unnatural amino acid mutagenesis in mapping ion channel function. *Curr. Opin. Neurobiol.* **13**, 264–270 (2003).
7. Hortin, G. & Boime, I. Applications of amino acid analogs for studying co- and posttranslational modifications of proteins. *Methods Enzymol.* **96**, 777–784 (1983).
8. Furter, R. Expansion of the genetic code: site-directed *p*-fluoro-phenylalanine incorporation in *Escherichia coli*. *Protein Sci.* **7**, 419–426 (1998).
9. Doring, V. *et al.* Enlarging the amino acid set of *Escherichia coli* by infiltration of the valine coding pathway. *Science* **292**, 501–504 (2001).
10. Kirshenbaum, K., Carrico, I. S. & Tirrell, D. A. Biosynthesis of proteins incorporating a versatile set of phenylalanine analogues. *Chembiochem* **3**, 235–237 (2002).
11. Benzer, S. & Champe, S. P. A change from nonsense to sense in the genetic code. *Proc. Natl Acad. Sci. USA* **48**, 1114–1121 (1962).
12. Garen, A. & Siddiqi, O. Suppression of mutations in the alkaline phosphatase structural cistron of *E. coli*. *Proc. Natl Acad. Sci. USA* **48**, 1121–1127 (1962).
13. Wang, L., Magliery, T. J., Liu, D. R. & Schultz, P. G. A new functional suppressor tRNA/aminoacyl-tRNA synthetase pair for the *in vivo* incorporation of unnatural amino acids into proteins. *J. Am. Chem. Soc.* **122**, 5010–5011 (2000).
14. Fechter, P., Rudinger-Thirion, J., Tukalo, M. & Giege, R. Major tyrosine identity determinants in *Methanococcus jannaschii* and *Saccharomyces cerevisiae* $tRNA^{Tyr}$ are conserved but expressed differently. *Eur. J. Biochem.* **268**, 761–767 (2001).
15. Steer, B. A. & Schimmel, P. Major anticodon-binding region missing from an archaebacterial tRNA synthetase. *J. Biol. Chem.* **274**, 35601–35606 (1999).
16. Jakubowski, H. & Goldman, E. Editing of errors in selection of amino acids for protein synthesis. *Microbiol. Rev.* **56**, 412–429 (1992).
17. Wang, L. & Schultz, P. G. A general approach for the generation of orthogonal tRNAs. *Chem. Biol.* **8**, 883–890 (2001).
18. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the genetic code of *Escherichia coli*. *Science* **292**, 498–500 (2001).
19. Kobayashi, T. *et al.* Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nature Struct. Biol.* **10**, 425–432 (2003).
20. Zhang, Y., Wang, L., Schultz, P. G. & Wilson, I. A. Crystal structures of apo wild-type *M. jannaschii* tyrosyl-tRNA synthetase (TyrRS) and an engineered TyrRS specific for *O*-methyl-L-tyrosine. *Protein Sci.* **14**, 1340–1349 (2005).
21. Ryu, Y. & Schultz, P. G. Efficient incorporation of unnatural amino acids into proteins in *Escherichia coli*. *Nature Methods* **3**, 263–265 (2006).
22. Anderson, J. C. & Schultz, P. G. Adaptation of an orthogonal archaeal leucyl-tRNA and synthetase pair for four-base, amber, and opal suppression. *Biochemistry* **42**, 9598–9608 (2003).
23. Anderson, J. C. *et al.* An expanded genetic code with a functional quadruplet codon. *Proc. Natl Acad. Sci. USA* **101**, 7566–7571 (2004).
24. Kowal, A. K., Kohrer, C. & RajBhandary, U. L. Twenty-first aminoacyl-tRNA synthetase–suppressor tRNA pairs for possible use in site-specific incorporation of amino acid analogues into proteins in eukaryotes and in eubacteria. *Proc. Natl Acad. Sci. USA* **98**, 2268–2273 (2001).
25. Santoro, S. W., Anderson, J. C., Lakshman, V. & Schultz, P. G. An archaebacteria-derived glutamyl-tRNA synthetase and tRNA pair for unnatural amino acid mutagenesis of proteins in *Escherichia coli*. *Nucleic Acids Res.* **31**, 6700–6709 (2003).
26. Mehl, R. A. *et al.* Generation of a bacterium with a 21 amino acid genetic code. *J. Am. Chem. Soc.* **125**, 935–939 (2003).
27. Chin, J. W. *et al.* An expanded eukaryotic genetic code. *Science* **301**, 964–967 (2003).
28. Wu, N., Deiters, A., Cropp, T. A., King, D. & Schultz, P. G. A genetically encoded photocaged amino acid. *J. Am. Chem. Soc.* **126**, 14306–14307 (2004).
29. Kiga, D. *et al.* An engineered *Escherichia coli* tyrosyl–tRNA synthetase for site-specific incorporation of an unnatural amino acid into proteins in eukaryotic translation and its application in a wheat germ cell-free system. *Proc. Natl Acad. Sci. USA* **99**, 9715–9720 (2002).
30. Sakamoto, K. *et al.* Site-specific incorporation of an unnatural amino acid into proteins in mammalian cells. *Nucleic Acids Res.* **30**, 4692–4699 (2002).
31. Zhang, Z. *et al.* Selective incorporation of 5-hydroxytryptophan into proteins in mammalian cells. *Proc. Natl Acad. Sci. USA* **101**, 8882–8887 (2004).
32. Hohsaka, T., Ashizuka, Y., Taira, H., Murakami, H. & Sisido, M. Incorporation of nonnatural amino acids into proteins by using various four-base codons in an *Escherichia coli in vitro* translation system. *Biochemistry* **40**, 11060–11064 (2001).
33. Anderson, J. C., Magliery, T. J. & Schultz, P. G. Exploring the limits of codon and anticodon size. *Chem. Biol.* **9**, 237–244 (2002).
34. Feinstein, S. I. & Altman, S. Context effects on nonsense codon suppression in *Escherichia coli*. *Genetics* **88**, 201–219 (1978).
35. Wang, L., Zhang, Z., Brock, A. & Schultz, P. G. Addition of the keto functional group to the genetic code of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **100**, 56–61 (2003).
36. Zhang, Z. *et al.* A new strategy for the site-specific modification of proteins *in vivo*. *Biochemistry* **42**, 6735–6746 (2003).
37. Chin, J. W. *et al.* Addition of *p*-azido-L-phenylalanine to the genetic code of *Escherichia coli*. *J. Am. Chem. Soc.* **124**, 9026–9027 (2002).
38. Deiters, A. *et al.* Adding amino acids with novel reactivity to the genetic code of *Saccharomyces cerevisiae*. *J. Am. Chem. Soc.* **125**, 11782–11783 (2003).
39. Deiters, A., Cropp, T. A., Summerer, D., Mukherji, M. & Schultz, P. G. Site-specific PEGylation of proteins containing unnatural amino acids. *Bioorg. Med. Chem. Lett.* **14**, 5743–5745 (2004).
40. Tsao, M. L., Tian, F. & Schultz, P. G. Selective Staudinger modification of proteins containing *p*-azidophenylalanine. *Chembiochem* **6**, 2147–2149 (2005).
41. Farrell, I. S., Toroney, R., Hazen, J. L., Mehl, R. A. & Chin, J. W. Photo-cross-linking interacting proteins with a genetically encoded benzophenone. *Nature Methods* **2**, 377–384 (2005).
42. Chin, J. W., Martin, A. B., King, D. S., Wang, L. & Schultz, P. G. Addition of a photocrosslinking amino acid to the genetic code of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 11020–11024 (2002).
43. Chin, J. W. & Schultz, P. G. *In vivo* photocrosslinking with unnatural amino acid mutagenesis. *Chembiochem* **3**, 1135–1137 (2002).
44. Kauer, J. C., Erickson-Viitanen, S., Wolfe, H. R. Jr & DeGrado, W. F. *p*-Benzoyl-L-phenylalanine, a new photoreactive amino acid. Photolabeling of calmodulin with a synthetic calmodulin-binding peptide. *J. Biol. Chem.* **261**, 10695–10700 (1986).
45. Schlieker, C. *et al.* Substrate recognition by the AAA+ chaperone ClpB. *Nature Struct. Mol. Biol.* **11**, 607–615 (2004).
46. Weibezahn, J. *et al.* Thermotolerance requires refolding of aggregated proteins by substrate translocation through the central pore of ClpB. *Cell* **119**, 653–665 (2004).
47. Hino, N. *et al.* Protein photo-cross-linking in mammalian cells by site-specific incorporation of a photoreactive amino acid. *Nature Methods* **2**, 201–206 (2005).
48. Deiters, A., Groff, D., Ryu, Y., Xie, J. & Schultz, P. G. A genetically encoded photocaged tyrosine. *Angew. Chem. Int. Edn Engl.* **45**, 2728–2731 (2006).
49. Bartels, E., Wassermann, N. H. & Erlanger, B. F. Photochromic activators of the acetylcholine receptor. *Proc. Natl Acad. Sci. USA* **68**, 1820–1823 (1971).
50. Bose, M., Groff, D., Xie, J., Brustad, E. & Schultz, P. G. The incorporation of a photoisomerizable amino acid into proteins in *E. coli*. *J. Am. Chem. Soc.* **128**, 388–389 (2006).

51. Rudd, P. M. & Dwek, R. A. Glycosylation: heterogeneity and the 3D structure of proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 1–100 (1997).
52. Zhang, Z. *et al.* A new strategy for the synthesis of glycoproteins. *Science* **303**, 371–373 (2004).
53. Xu, R. *et al.* Site-specific incorporation of the mucin-type *N*-acetylgalactosamine-α-*O*-threonine into protein in *Escherichia coli*. *J. Am. Chem. Soc.* **126**, 15654–15655 (2004).
54. Lu, W., Gong, D., Bar-Sagi, D. & Cole, P. A. Site-specific incorporation of a phosphotyrosine mimetic reveals a role for tyrosine phosphorylation of SHP-2 in cell signaling. *Mol. Cell* **8**, 759–769 (2001).
55. Xie, J. *Adding Unnatural Amino Acids to the Genetic Repertoire.* Thesis, Scripps Research Institute, La Jolla (2006).
56. Wang, J., Xie, J. & Schultz, P. G. A genetically encoded fluorescent amino acid. *J. Am. Chem. Soc.* **128**, 8738–8739 (2006).
57. Summerer, D. *et al.* A genetically encoded fluorescent amino acid. *Proc. Natl Acad. Sci. USA* **103**, 9785–9789 (2006).
58. Tsao, M. L., Summerer, D., Ryu, Y. & Schultz, P. G. The genetic incorporation of a distance probe into proteins in *Escherichia coli*. *J. Am. Chem. Soc.* **128**, 4572–4573 (2006).
59. Xie, J. *et al.* The site-specific incorporation of *p*-iodo-L-phenylalanine into proteins for structure determination. *Nature Biotechnol.* **22**, 1297–1301 (2004).
60. Deiters, A., Geierstanger, B. H. & Schultz, P. G. Site-specific *in vivo* labeling of proteins for NMR studies. *Chembiochem* **6**, 55–58 (2005).
61. Suydam, I. T. & Boxer, S. G. Vibrational Stark effects calibrate the sensitivity of vibrational probes for electric fields in proteins. *Biochemistry* **42**, 12050–12055 (2003).
62. Alfonta, L., Zhang, Z., Uryu, S., Loo, J. A. & Schultz, P. G. Site-specific incorporation of a redox-active amino acid into proteins. *J. Am. Chem. Soc.* **125**, 14662–14663 (2003).
63. Wang, L., Xie, J., Deniz, A. A. & Schultz, P. G. Unnatural amino acid mutagenesis of green fluorescent protein. *J. Org. Chem.* **68**, 174–176 (2003).
64. Wang, L., Brock, A. & Schultz, P. G. Adding L-3-(2-naphthyl)alanine to the genetic code of *E. coli*. *J. Am. Chem. Soc.* **124**, 1836–1837 (2002).
65. Turner, J. M., Graziano, J., Spraggon, G. & Schultz, P. G. Structural characterization of a *p*-acetylphenylalanyl aminoacyl-tRNA synthetase. *J. Am. Chem. Soc.* **127**, 14976–14977 (2005).
66. Turner, J. M., Graziano, J., Spraggon, G. & Schultz, P. G. Structural plasticity of an aminoacyl-tRNA synthetase active site. *Proc. Natl Acad. Sci. USA* **103**, 6483–6488 (2006).
67. Tian, F., Tsao, M. L. & Schultz, P. G. A phage display system with unnatural amino acids. *J. Am. Chem. Soc.* **126**, 15962–15963 (2004).
68. Castelli, D. D., Lovera, E., Ascenzi, P. & Fasano, M. Unfolding of the loggerhead sea turtle (*Caretta caretta*) myoglobin: a ^{1}H-NMR and electronic absorbance study. *Protein Sci.* **11**, 2273–2278 (2002).

Acknowledgements
Work in the laboratory of P.G.S. is supported by the National Institutes of Health, the United States Department of Energy and the Skaggs Institute for Chemical Biology.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
UniProtKB: http://ca.expasy.org/sprot
ClpB | Gal4 | GRB2 | LamB | STAT1

FURTHER INFORMATION
Peter G. Schultz's homepage:
http://schultz.scripps.edu
Access to this links box is available online.

OPINION

SUN-domain proteins: 'Velcro' that links the nucleoskeleton to the cytoskeleton

Yonatan B. Tzur, Katherine L. Wilson and Yosef Gruenbaum

Abstract | The novel SUN-domain family of nuclear envelope proteins interacts with various KASH-domain partners to form SUN-domain-dependent 'bridges' across the inner and outer nuclear membranes. These bridges physically connect the nucleus to every major component of the cytoskeleton. SUN-domain proteins have diverse roles in nuclear positioning, centrosome localization, germ-cell development, telomere positioning and apoptosis. By serving both as mechanical adaptors and nuclear envelope receptors, we propose that SUN-domain proteins connect cytoplasmic and nucleoplasmic activities.

For a discussion of the SUN-domain (Sad1 and UNC-84 homology domain) family of proteins, we must first introduce the nuclear envelope. In all eukaryotic cells, the nuclear envelope separates the nucleus from the cytoplasm. The nuclear envelope is composed of an outer nuclear membrane (ONM) and an inner nuclear membrane (INM). The two membranes join at nuclear pore complexes, which control the traffic of macromolecules between the nucleoplasm and the cytoplasm. In metazoans, the nuclear pore complexes and the INM are anchored to a structural network of lamin filaments[1] (BOX 1). Lamin polymers confer mechanical strength to the nucleus[2,3]. Many nuclear membrane and nucleoplasmic proteins interact with lamins[4], and lamin-associated protein complexes support a broad range of functions[4] (BOX 1). At least 80 unique integral membrane proteins were found to localize at the nuclear envelope in mammalian cells[5]. It is assumed that most of these proteins interact directly or indirectly with lamins[4]. Among these nuclear membrane proteins are three families, each of which is characterized by a distinct motif (specifically, LEM, SUN or KASH). LEM-domain proteins (named after LAP2, emerin and MAN1) are reviewed elsewhere[4]. We will discuss proteins that contain the SUN domain and their partners, many of which contain the KASH (named after Klarsicht, ANC-1 and SYNE1 homology) domain[6–8].

Most (but not all) SUN-domain proteins are localized at the INM. The situation is more complicated for KASH-domain proteins, most individual isoforms of which are localized on either the ONM or the INM[8,9]. However, some isoforms lack the KASH domain, do not localize to the nuclear envelope and are instead proposed to tether other organelles to the cytoskeleton[8]. Several INM-localized SUN-domain and KASH-domain proteins can interact with lamins, and in certain cases this interaction is required for their nuclear envelope localization (see below).

This perspective will focus primarily on SUN-domain proteins: we will depict their structural organization (known and hypothetical) in complexes that traverse the nuclear envelope, and discuss their roles in nuclear positioning, centrosome attachment to the ONM, links to the cytoskeleton, and telomere positioning during meiosis. We will propose a model in which SUN-domain proteins serve as mechanical 'Velcro' to interconnect the cytoskeleton and nucleoskeleton, and suggest that SUN-domain proteins have further, non-mechanical roles as specialized nuclear envelope receptors.

SUN-domain and KASH-domain proteins

Both the SUN domain and regions that are upstream of the SUN domain interact directly with the KASH domain of KASH-domain proteins[10,11]. This interaction is required for the cellular functions of both types of protein.

Domain organization of SUN-domain proteins. Malone and colleagues[12] coined the term SUN domain when they discovered a motif of ~120 residues in the C terminus of the *Caenorhabditis elegans* UNC-84 protein. This protein has significant homology to a region in the *Schizosaccharomyces pombe* Sad1 protein and several uncharacterized mammalian proteins. Genome-database