Toward the experimental codon reassignment *in vivo*: protein building with an expanded amino acid repertoire

NEDILJKO BUDISA,[1] CAROLINE MINKS, STEFAN ALEFELDER, WALTRAUD WENGER, FUMIN DONG, LUIS MORODER, AND ROBERT HUBER

Max Planck Institut für Biochemie, D-82152 Martinsried, Germany

ABSTRACT **The high precision and fidelity of the genetic message transmission are ensured by numerous proofreading steps, from DNA replication and transcription to protein translation. The key event for translational fidelity is the proper codon assignment for 20 canonical amino acids. An experimental codon reassignment is possible for noncanonical amino acids *in vivo* using artificially constructed expression hosts under efficient selective pressure. However, such amino acids may interfere with the cellular metabolism and thus do not belong to the 'first' or 'restricted' part of the universal code, but rather to a second or 'relaxed' part, which is limited mainly by the downstream proofreading in the natural translational machinery. Correspondingly, not all possible α-amino acids can be introduced into proteins. The aim of this study is to discuss biological and evolutionary constraints on possible candidates for this second coding level of the universal code. Engineering of such a 'second' code is expected to have great academic as well as practical impact, ranging from protein folding studies to biomedicine.—Budisa, N., Minks, C., Alefelder, S., Wenger, W., Dong, F., Moroder, L., Huber, R. Toward the experimental codon reassignment *in vivo*: protein building with an expanded amino acid repertoire. *FASEB J.* 13, 41–51 (1999)**

Key Words: genetic code · protein folding · translation

TERMINOLOGY

The term 'genetic code' refers to the correspondence between nucleotide sequence and amino acid sequence (**Fig. 1**). The universality of the genetic code stems from the nearly identical ('canonical') assignment of amino acids to triplet codons in all living organisms, thus affirming their common origin (1). In general, these amino acids must allow the folding of polypeptide chains into specific, compact, and balanced conformations able to carry out sophisticated chemical operations *in vivo* (2).

Most of the functional requirements of proteins can be fulfilled with the standard set of 20 α-amino acids. Additional functional diversity is achieved via posttranslational modifications or cotranslational incorporation of amino acids into proteins that do not belong to the standard set. Such α-amino acids [e.g., selenocysteine (3) or lanthionine (4), etc.] are also natural, proteinogenic, or common, and even 'canonical' in the context of their appearance.

For example, amino acids like selenocysteine (SeCys) or formyl-methionine (fMet), which are deciphered in special cases in a context-dependent manner (as result of local redefinition of codon meaning), are defined as 'special canonical amino acids' (**Scheme 1**). Furthermore, selenocysteine is incorporated as a response to the UGA codon in a few *Escherichia coli* proteins by a particular mechanism that keeps an internal UGA codon separated from the terminal UGA stop codons (3). Moreover, there are other special examples where context-dependent deciphering of genetic code also works for canonical amino acids, such as nonsense suppression phenomena (5, 6). In the artificial laboratory conditions of *in vivo* protein expression, it should be possible to extend the number of amino acids encoded by the universal code. A new amino acids coded in this noncanonical manner also belong to the universal code.

Because of the lack of a broad consensus about the nomenclature of α-amino acid outside of the canonical 20, we propose the term 'noncanonical amino acids' (Scheme 1) for those α-amino acids that can be successfully integrated translationally into proteins in a residue-specific manner, and the term 'special noncanonical amino acids' for those whose possible translational introduction into proteins *in vivo* is context dependent. Other terms, including isostere, analog, or homologue, can be conveniently used for both noncanonical (e.g., seleno-methionine as isosteric analog of methionine) and canonical amino acids (e.g., cysteine as isosteric analog of serine, and vice versa). Naturally occurring α-amino ac-

[1] Correspondence: Max Planck Institut für Biochemie, Abteilung Strukturforschung, Am Klopferspitz 18A, D-82152 Martinsried, Germany. E-mail: budisa@biochem.mpg.de

0892-6638/99/0013-0041/$02.25 © FASEB

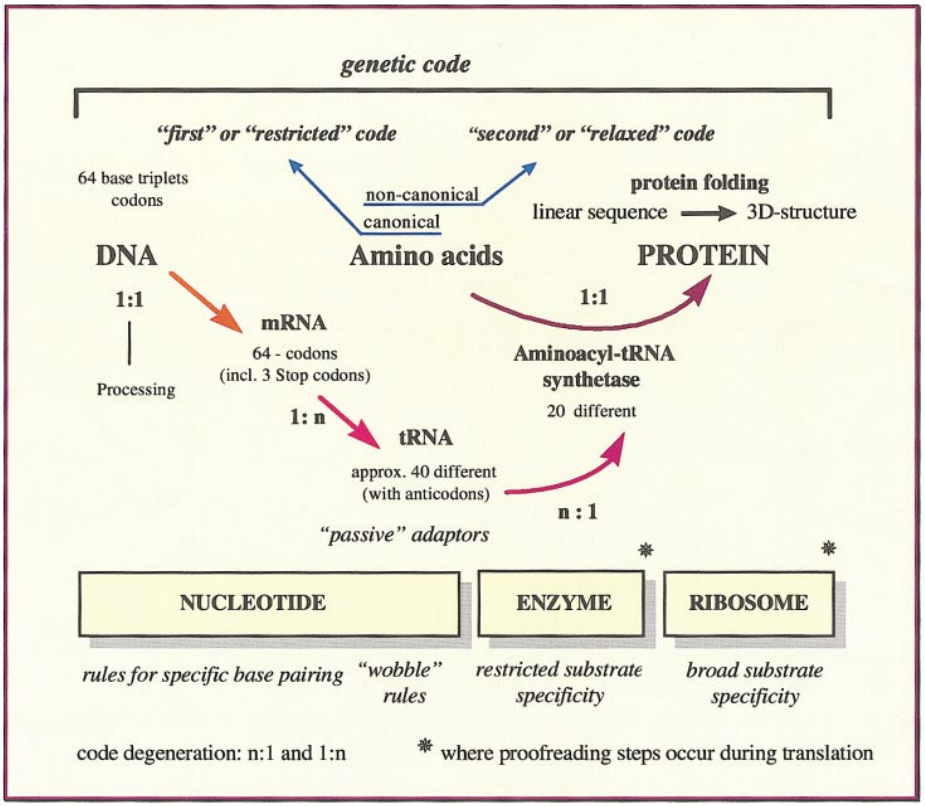

Figure 1. The genetic code is based on the complicated relations between base triplets in DNA or mRNA linear sequence and amino acids integrated into protein structures. The code is degenerated (n:1, 1:n), recognized, and translated by aminoacyl-tRNA synthetases through specific pairing of the appropriate tRNA with the related amino acid. The high precision of the genetic message transmission is further ensured by proofreading steps from DNA replication and transcription to the protein translation. The translational process includes activation of amino acids, proofreading during activation, synthesis on ribosome, and proofreading during synthesis. All these processes are mutually interdependent and intrinsically coupled with the polypeptide chain folding into a specific 3-dimensional structure, in this way allowing translation of the original genetic message written in DNA sequence into biological activity.

ids that arise from posttranslational modifications or from secondary metabolism are named 'special biogenic amino acids' (Scheme 1). These amino acids were identified as intermediates of metabolic pathways or as products of the metabolism or detoxification of foreign compounds.

AIM OF THIS STUDY

The precise transmission of the nucleotide genetic messages into proteins is partially fulfilled by stringent stereochemical substrate requirements of the translational machinery. Much effort in recent years has been devoted to the expansion of a pool of amino acids that can be utilized by the translational machinery for an efficient protein biosynthesis *in vivo*. For academic as well as practical purposes, it is useful to define factors that permit as well as limit a simple, practical, and efficient codon reassignment in the framework of the existing universal genetic code. The possibilities and prospects of such reassignment are discussed here in the context of the most recent research performed under *in vivo* conditions for protein expression with an expanded amino acid repertoire. It is expected that the expression of protein mutants with noncanonical amino acids should allow delineation of the evolutionary balance between stability and recycling as well as stability and functionality of expressed proteins.

SCHEME 1. *Proposed nomenclature*

α-Amino acids*		Corresponding processes
Canonical	Noncanonical	—Translational codon-dependent incorporation
Special canonical	Special noncanonical**	—Translational codon + context-dependent incorporation
Special biogenic	—	—Posttranslational modifications
		—Metabolic precursors and intermediates
—Vital for cellular viability	—Toxic effects on cellular viability	—*In vivo* effects and natural abundance
—Abundant in nature	—Appears in smaller amounts	

* A virtually countless number of α-amino acids can be or are prepared in the laboratory. On the other hand, only a tiny portion of them can be *in vivo* cotranslationally introduced into proteins, whereas most can be built into peptides by peptide synthesis protocols. Although total chemical synthesis of proteins may be an interesting direction for future research, it is beyond the scope of this study, which is concerned mainly with *in vivo* protein synthesis. ** Site-directed introduction of special noncanonical amino acids (for review, see refs 6, 7) occurs in the context of high suppression of specific stop codons and in the presence of chemically acylated (charged) tRNAs. Current attempts to transfer this *in vitro* method to *in vivo* conditions are described in ref 31.

FIDELITY AND EDITING IN THE TRANSLATION FROM AMINOACYLATION TO RIBOSOME BIOSYNTHESIS

The accuracy of the translational process depends on the precision of two successive independent recognition events, that of *1*) amino acids with tRNAs, and *2*) charged tRNAs with ribosome-linked mRNAs. The initial step in protein synthesis is covalent linkage of defined amino acids to cognate tRNAs. This process is mediated by aminoacyl-tRNA synthetases, enzymes that act as 'interpreters' of the genetic code through the aminoacylation reaction, where tRNAs are charged with the 'correct' amino acids. The charged tRNAs act as shuttles for the delivery of the amino acids to the ribosome. On the ribosome, the genetic code is decoded by RNA-RNA base pairing interactions between the tRNA anticodon nucleotide triplet and the complementary mRNA codon (7).

The important feature of the process of protein translation is a very low rate of misincorporation at the amino acids level. The overall error rate in protein synthesis is less than 3×10^{-4} (8) whereas the frequency of errors in the aminoacylation reaction are probably even lower than this (9). Such accuracy is achieved by a precise assignment of specific codons to specific amino acids. The question of how this accuracy is achieved has been the subject of extensive research in recent years.

At the level of the aminoacylation reaction, the error frequency is composed of two factors, including the relative efficiency of the enzyme for different amino acids as substrates and the relative concentration of competing substrates in the cytosol (10). Several aminoacyl-tRNA synthetases achieve their high fidelity by an editing or proofreading mechanism (11). Although ribosomes exhibit a broad substrate specificity, proofreading also occurs at this level of protein biosynthesis; often noncognate complexes, such as D-amino acid-charged tRNAs, are efficiently hydrolysed prior to chain elongation (5).

Finally, protein folding is an integral part of this process (Fig. 1). It has been hypothesized that the amino acid sequence determines the 3-dimensional protein structure by means of a stereochemical code (12), whereby a set of specific rules discriminates between native conformation and other conceivable chain folds (13). Although it is not excluded that folding might occur posttranslationally, there is a growing body of evidence that ribosomes themselves may mediate folding of the nascent proteins during their synthesis. According to this scenario, protein folding of nascent domains is completed before hydrolysis of the peptidyl-tRNA, which occurs during codon-directed termination and release of the protein from the ribosome (14).

In conclusion, editing is especially restrictive toward: *1*) metabolic precursors and intermediates, e.g., homocysteine as precursor of methionine in its biosynthesis (**Fig. 2**), *2*) similar compounds relatively abundant in cytoplasm, e.g., leucine and valine, which differ for only one $-CH_2-$ group, *3*) amino acids that sterically and chemically differ significantly from canonical counterparts, and *4*) amino acids that act as inhibitors of protein translation, e.g., selenaproline, which can be activated and charged on $tRNA^{Pro}$ but inhibits chain elongation on the ribosome (15, 16).

'RESTRICTED' VS. 'RELAXED' GENETIC CODE

The enzymes of the proofreading machinery have not evolved to exclude or to 'edit' the specific features of some sterically or chemically similar noncanonical amino acids and allow for their efficient introduction into proteins during biosynthesis. Thus, a rather large number of noncanonical amino acids that resemble canonical counterparts by shape, size,

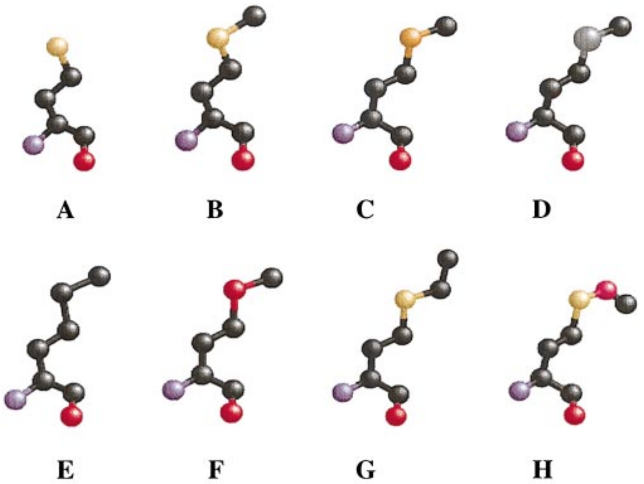

Figure 2. The canonical amino acid methionine and some methionine-like noncanonical amino acids. The translational cellular machinery regularly recognizes methionine (*B*) as natural substrate. In addition, residue-specific replacement is possible for selenomethionine (*C*), telluromethionine (*D*), norleucine (*E*), and ethionine (*G*). Homocysteine (*A*) is the metabolic precursor of methionine and therefore efficiently edited during translation. For other methionine-like, noncanonical amino acids, bioincorporation is not possible even under conditions of strong selective pressure. Methoxinine, with oxygen at the δ-position (*F*), although sterically similar to methionine, differs greatly in its physico-chemical properties (it is more hydrophilic than methionine; Pauling electronegativity for oxygen is 3.44, for sulfur it is 2.58) and therefore it is not introduced into proteins. The ethionine-like mercury containing amino acid methyl-mercury-homocysteine (*H*) is highly poisonous for living cells and to the translational machinery *per se* due to the complex chemical reactivity of mercury.

and chemical properties have been incorporated successfully *in vivo* into proteins. For example, we have recently reported the bioincorporation of the methionine-like noncanonical amino acids ethionine, norleucine, seleno-, and telluromethionine into human recombinant annexin V (Fig. 2).

A target of the editing activity of the methionyl-tRNA synthetase is the methionine metabolic precursor homocysteine. Consequently, methionine-like amino acids smaller in size than methionine per se are edited, whereas norleucine is on the border of such size based exclusion (11). In the absence of an efficient selective pressure, methionine is preferentially activated and incorporated into living cells (17); in the *in vitro* aminoacylation reaction, methionine is also preferably activated by methionyl-tRNA synthetase (11). Although some of these methionine-related noncanonical amino acids are incorporated into proteins (Fig. 2), their actual presence would be highly toxic for all cells; for example, norleucine inhibits cell growth and its incorporation into cellular proteins would be lethal for the living cells, thus imposing stringent evolutionary barriers. Although norleucine is abundant in carbonaceous Murchison meteorites and apparently represents a product of abiotic synthesis in greater amounts than most of the canonical amino acids (18), its toxicity for present living cells excludes this amino acid from the universal code. The 'restricted' part of the universal code encodes only for those amino acids that are optimally integrated into the metabolic chemistry of the living cells and allow their reproduction, growth, and differentiation. On the other hand, most of the noncanonical amino acids in this context are 'xeno-compounds'. These amino acids can be integrated into proteins under *in vivo* conditions only in the presence of a mechanism for efficient elimination of toxic effects for the cellular metabolism. This can be achieved by using artificially constructed hosts with highly sophisticated and fully controllable recombinant protein expression systems. Correspondingly, these noncanonical amino acids belong to the relaxed or second part of the universal code.

Among the methionine-like noncanonical amino acids, we find a limited set of amino acids for which bioincorporation is possible (Fig. 2). Their common feature in translation is not only that they escape the editing mechanisms, but they also permit effective protein folding such that stability and functionality are retained. In this context, we have recently studied folding parameters of methionine-containing (wild-type form) and methionine-substituted annexin V protein variants (19). Methionine translational integration was found to generate a protein structure that is characterized by a higher cooperative behavior as judged from the slopes of unfolding curves in comparison with the protein variants that contain other similar, but physicochemically different, noncanonical amino acids (19).

It has been postulated that protein sequences that have been optimized through evolution by using the canonical amino acids as building blocks become functionally inferior once they contain noncanonical amino acids (20). However, a priori it cannot be excluded that noncanonical amino acids might lead to improved stability or functional properties of proteins. Recently, we discovered that residue-specific replacement of all phenylalanines with *ortho*-fluorophenylalanines in human recombinant annexin V results in higher stability and folding cooperativity (C. Minks, R. Huber, L. Moroder, and N. Budisa, unpublished results). Similarly, replacement of the active-site histidine with 3-triazolyl-alanine in porcine pancreatic phospholipase yielded an enzyme active at more acidic pH values (21). These examples confirm that improved stabilities and cooperativities as well as enzymatic activities can be obtained by integration of noncanonical amino acids into individual proteins. However, cellular lethality of such amino acids (as discussed above), when integrated into other cellular proteins, prohibits their usage in the restricted part of the universal code. We postulate that the universality of the code is not limited to its restricted ('first') part, but consists also of a relaxed or second part, which can be experimentally assessed.

TRIPLET CODING EXPANSION: ENGINEERING A SECOND CODE

Evolutionary selection pressure has directed translation toward producing optimally folded and functional proteins. This is achieved by folding patterns that generate 'active sites' via precise 3-dimensional arrangement of functional groups (2). One of the most important biological constraints for such protein production is the proper codon assignment optimized during evolution. On the other hand, this highly restricted codon assignment can be bypassed under artificial conditions. After exclusion of toxic effects caused by bioincorporation of various noncanonical amino acids, it might be possible to disclose the rules or determinants that allow for an experimental extension of the coding properties of particular triplets. Such determinants must be closely related to general principles that govern properly balanced protein stabilities and functionalities. However, other factors should not be neglected, such as the uniqueness and specificity of the folded structure, its marginal stability, the packing of a protein interior, and integration of standard 20 amino acids into processes other than protein building (metabolic precursors, intermediates, etc.) (20, 22).

As discussed, some limits for possible candidates for this additional coding level (second or relaxed

TABLE 1. *Examples of codon assignment under natural conditions (canonical amino acids) and reassignment (non-canonical amino acids) under strong selective pressure produced in the artificially designed expression systems in vivo*[a]

	Amino acids assigned		
Codons	Canonical	Noncanonical	Reference
AUG	Methionine	Selenomethionine, telluromethionine, ethionine, norleucine	17, 45
UGG	Tryptophan	4-Fluorotryptophan, 5-fluorotryptophan, 6-fluorotryptophan, 7-azatryptophan, 5-hydroxytryptophan	36, 39, 50
UU^{U}_{C}	Phenylalanine	*o*-Fluorophenylalanine, *p*-fluorophenylalanine, *m*-fluorophenylalanine, thienylalanine	34, 35, Minks et al., unpublished results
UA^{U}_{C}	Tyrosine	*m*-Fluorotyrosine	38, Minks et al., unpublished results
$CC^{\substack{U\\C}}_{\substack{A\\G}}$	Proline	Thioproline	16
UG^{U}_{C}	Cysteine	Selenocysteine	49

[a] Expansion of the coding properties (relaxed or second code) for each triplet or family of triplets is possible only for those amino acids that can be introduced into stable and functional proteins in a residue-specific manner. Position of selenocysteine is unique. It is a special canonical amino acid [response to UGA codon (25)]; since its *in vivo* efficient residue-specific bioincorporation is possible (response to cysteine UGU and UGG codons) (50), it is also a noncanonical amino acid. Successful introductions are referenced according to the criteria described in this study.

code) must exist, and it might be possible to define them experimentally. The study of such limitations would provide a pool of noncanonical amino acids that are translationally integrated in a residue-specific manner. Methionine and methionine-like noncanonical amino acids (Fig. 2) represent one example of new codon reassignments. The AUG codon at the level of the first code assigns methionine as canonical amino acid, whereas at the level of the second code it assigns for the sterically similar, but physicochemically different residues such as selenomethionine, telluromethionine, norleucine, and ethionine as noncanonical amino acids (**Table 1**). In addition to AUG, it should be possible to define such a list for each codon or codon families. Furthermore, context-dependent incorporations *in vivo* should result in a list of special noncanonical amino acids.

POSSIBLE EXPERIMENTAL CRITERIA

The question remains as to which experimental criteria might be used for adding a particular α-amino acid to the list of noncanonical amino acids or the special noncanonical amino acids at the level of the second code. We propose three simple but strict experimental requirements for such codon expansion: *1)* noncanonical amino acid should be directly incorporated into various target proteins *in vivo* and the resulting protein variants isolated; *2*) bioincorporation must be confirmed by at least two different analytical techniques, e.g., amino acid analysis and mass spectrometry (**Fig. 3**), and *3*) the protein mutants should be structurally and functionally analyzed to assess whether the variants are stably folded and whether activity of the protein has been altered.

CODON REASSIGNMENT IN THE LIVING CELLS

A codon assignment in living cells relies on the high selectivity of aminoacyl-tRNA synthetases in their recognition of both the amino acid and the related tRNA. The classical dogma is that there is one aminoacyl-tRNA synthetase in the cell for each canonical amino acid (Fig. 1). However, at least two well-documented examples demonstrate how nature can develop a new or alternative pathway for the incorporation of new amino acids into proteins without creating a new aminoacyl-tRNA synthetase.

In gram-positive eubacteria and organelles of eukaryotes that lack glutaminyl-tRNA synthetase, the glutamine-charged tRNA is generated cotranslationally. tRNA^{Gln} is charged ('mischarged') with glutamic acid by glutamyl-tRNA synthetase and then tRNA-bound glutamic acid is converted to tRNA-bound glutamine by the specific enzyme tRNA^{Gln}-aminotransferase (23). A second and more complicated example is represented by the special canonical amino acid selenocysteine. It seems reasonable to expect that selenocysteine incorporation might result from a post-translational modification of cysteine rather than its incorporation into the growing polypeptide chain from a charged tRNA during chain elongation (24). But the mechanism for selenocysteine introduction in *E. coli* proteins is cotranslational and requires the presence of a UGA stop codon as well as an adjacent,

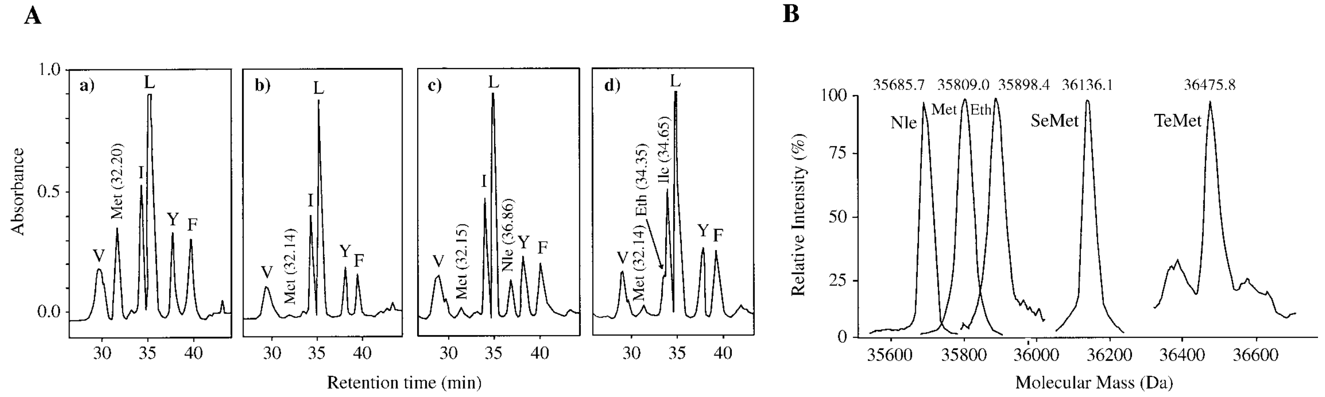

Figure 3. Analytical proof for the *in vivo* translational introduction of methionine-like noncanonical amino acids norleucine (Nle), ethionine (Eth) selenomethionine (SeMet), and telluromethionine (TeMet) into human recombinant annexin V. *A*) Amino acid analysis. Portions of the amino acid composition chromatograms for native and analog-containing human recombinant annexin V. About 5 μg of protein sample was hydrolyzed by the trifluoroacetic acid/hydrochloric acid vapour phase method (47). Absorbance peaks of ninhydrine derivatives of the amino acids are marked by a one-letter code (L, leucine; V, valine; Y, tyrosine, and F, phenylalanine). *a*) Native annexin V; *b*) SeMet introduction leads to nearly complete disappearance of the native methionine peak because SeMet is hydrolyzed under the standard reaction conditions; the same holds true for TeMet-containing protein; *c*) Nle-containing annexin V; *d*) Eth-annexin V; note that Eth appears in the chromatogram as a shoulder of the isoleucine peak. *B*) Electrospray ionization mass spectrometric (ESI-MS) analysis (48). Deconvoluted spectra from five separate measurements are superimposed at the same mass scale. Observed average masses are within the error range ±3.8 Da and correlate well with the expected masses, with the exception of Eth (deficit: 8–12 Da). See refs 17, 37 for more experimental details.

specific mRNA hairpin structure. Furthermore, a specific tRNA containing the UCA anticodon requires acylation with serine by seryl-tRNA-synthetase, followed by enzymatic conversion into selenocysteine-tRNA, which decodes the UGA codon. Finally, the presence of a special elongation factor *SelB* that recognizes both mRNA hairpin and selenocysteine-tRNA at the ribosome is required (25).

The question arises as to why the selenocysteine context-dependent integration into proteins occurs cotranslationally. The most rational explanation relies on the significant chemical differences between cysteine and its isosteric selenocysteine. At physiological pH 7, the cysteine thiol function is not deprotonated, whereas selenocysteine exists almost exclusively as the selenide ion (26). Since the redox potential of selenocysteine incorporated into model peptides (−382 mV) is much more reducing than that of related cysteine-peptides (−180 mV) (27), its circulating form in the cells would mainly be selenocystine. In the crystal structure of the selenocysteine-enzyme gluthatione peroxidase, the selenocysteine is located in a flat depression caged with aromatic and hydrophobic amino acids that sequester the residue from the bulk solvent (28). Therefore, the complex chemistry of selenium that requires protection form the bulk solvent, i.e., water exclusion, represents the most probable cause for the cotranslational selenocysteine bioincorporation that occurs even in evolutionary advanced cells. In other words, an efficient system has evolved that allows selenocysteine protection during its translational integration into proteins.

Generally, the most straightforward codon reassignment in nature among different organisms occurs evolutionary from one canonical amino acid to another. An example is the case of vertebrate mitochondria, which read AUA as methionine, whereas in the vertebrate cytosol this codon is assigned to isoleucine. An exceptional and unique example is the UGA codon that reads for stop, selenocysteine, tryptophan, and cysteine (18). Such diversity in the assignment of certain codons in the universal code suggests the possibility of experimental accommodation of additional amino acids into proteins (24).

EXPERIMENTAL CODON REASSIGNMENT BY ELIMINATION OF THE AMINOACYLATION PROOFREADING STEP

The ribosome translates mRNA messages into proteins with very low levels of errors due to misincorporations. At the same time, the translational machinery has rather broad substrate specificity, since it accommodates all types of canonical amino acids. The same holds true for many α-amino acids outside the set of the canonical 20. Thus, a promising approach for introduction of noncanonical amino acids is to avoid or bypass the aminoacylation step from the translation process.

In this way, much effort has been devoted in recent years to develop new and efficient methods of site-directed incorporation of various amino acids by recruiting the amber terminator codon UAG (6). The main advantage of this method is that the proofreading step during the amino acid activation is efficiently bypassed. However, the proofreading during ribosomal translation *in vitro* is again a serious limitation. Nevertheless, a variety of amino acids have been successfully incorporated into proteins by this procedure, making this *in vitro* suppression method (5) an excellent tool for probing ribosomal editing tolerance.

Many special noncanonical amino acids have interesting functions within proteins, acting as cages, sensors, or improved nucleophiles, electron donors, or acceptors. They have certainly proved useful for many academic and applicative purposes. Thus, transfer of this strictly *in vitro* procedure to an *in vivo* translational expression in artificially constructed host systems is desirable. However, the specific prerequisites for the context-dependent selenocysteine bioincorporation in the living cells, as discussed above, illustrate the challenges this approach will face. Thus, any attempt to tackle this problem should consider the translational process as the place where two worlds interact in a tuned and optimized way–namely, the protein and RNA world, where even subtle deviations could easily result in dangerous or fatal effects. One example of such an induced lethality is tRNA synthetase-induced cell death (29). Despite the anticipated difficulties, generation and optimization of such a system even for a single special noncanonical amino acid would be a substantial achievement (30, 31).

RELAXATION OF SUBSTRATE SPECIFICITIES OF AMINOACYL-tRNA SYNTHETASES

Another promising approach for exploiting the broad ribosome substrate specificity would be to change or relax the substrate specificity of the aminoacyl-tRNA synthetases. The first report in this direction demonstrated that it is possible to introduce amino acids like *para*-chloro-phenylalanine and *para*-bromo-phenylalanine into a recombinant luciferase that contains 29 phenylalanine codons (32). Both *in vitro* and *in vivo* incorporation of these amino acids does not affect the efficiency of translation of the full-length protein, but does result in more than 99% reduced protein activity. Even addition of various chaperones did not help to recover the activity. The lesson is obvious: even if it is possible to force the translation machinery to accommodate such amino acids, one cannot force proteins to fold correctly with such bulky atoms mainly within the hydrophobic core. This observation clearly shows the importance of the naturally occurring, efficient proofreading systems that prevent incorporation of similar compounds such as brominated or iodinated aromatic amino acids.

We recently tried to replace 12 mainly core-located tyrosine residues with *meta*-fluoro- and *meta*-chloro-tyrosine in various proteins by applying the strong selective pressure procedure, without attempts to relax the substrate specificity of tyrosyl-tRNA synthetase. Whereas the fluorinated amino acid could be easily introduced, bioincorporation of the chloro-compound occurs only to very low levels, as shown in **Fig. 4**. This clearly indicates that activation, loading on ribosomes, chain elongation, and even protein folding with *meta*-chloro-tyrosine take place to some extent in the presence of the natural substrate tyrosine. Its further gradual depletion obviously then leads to the cessation of the chain elongation process (C. Minks, R. Huber, L. Moroder, and N. Budisa, unpublished results). Although unknown factors may intervene with incorporation of this chlorinated amino acid at the level of ribosomal synthesis, perhaps the major deterrent to successful incorporation is the bulkiness and the difficult accommodation of this residue in the folded protein. On the other hand, sterical bulkiness cannot be regarded as the sole stumbling block to successful translational integration of various noncanonical amino acids (see Fig. 2).

This data suggest that all proofreading steps in such translational process have been optimized to allow for proper folding of the target protein; thus, residues introducing bulkiness like bromo and iodo substitutions in aromatic amino acids are recognized by the protein translation machinery and, consequently, even chain elongation is prevented from occurring (Fig. 4). Therefore, posttranslational modifications achieved by simple enzymatic reactions at surface-exposed residues are the best choice in this context. In fact, *in vivo* posttranslational halogenation is particularly well documented for iodination of tyrosine residues by thyroglobulin in wide variety of organisms, from tunicates to mammals (33). These amino acids cannot be listed as noncanonical amino acids, but rather as special biogenic amino acids (Scheme 1) generated post festum of the translational process as integral part of the protein functionality.

The consequences and lessons from such considerations are far-reaching. Even if substrate recognition of aminoacyl-tRNA synthetases and ribosomes is not necessarily strict in discriminating canonical from noncanonical amino acids, a firm boundary apparently exists between 'allowable' and nonpermissive amino acids at a structural level. The reasons are obvious, since the translational processes in living cells consist of the activation of amino acids, proofreading during the activation step, synthesis on ribosome, and proofreading during synthesis; these steps are mutually interdependent and intrinsically coupled with the proper folding of the resulting protein (Fig. 1). This might be the key determinant responsible for realization of the genetic message written in the DNA sequence in terms of sufficient and balanced protein stability and functionality. Similar

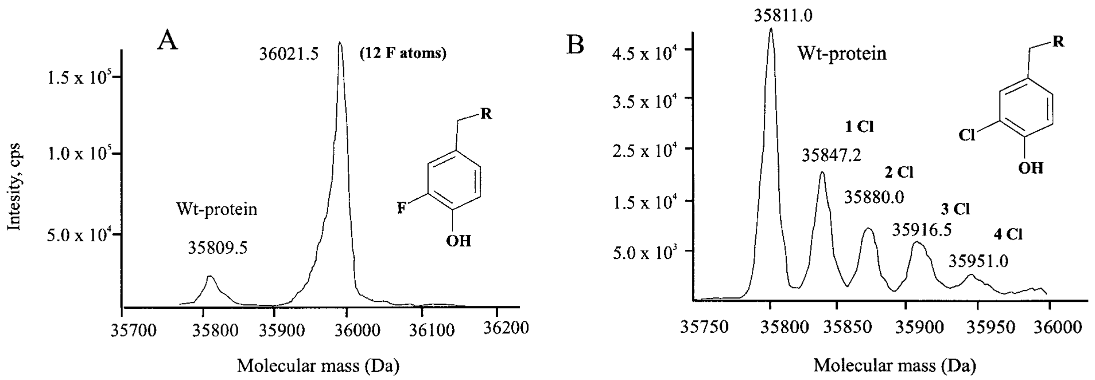

Figure 4. Electrospray ionisation mass spectrometric (ESI-MS) analyses of *meta*-fluoro- and *meta*-chloro-tyrosine introduction into human recombinant annexin V, which contains 12 codons for tyrosine. *A*) Small peak corresponds to the tyrosine-containing wt protein form (35809.5 Da); a dominant peak reveals the mass (36021.5 Da) nearly identical to the expected mass (additional 12 F atoms). *B*) Attempt for *meta*-chloro-tyrosine bioincorporation under the same experimental conditions. Note that wt peak intensity corresponds approximately to the intensity of the wt peak in panel *A*, indicating very low amounts of the protein, often below background expression (for experimental details, see Minks et al., unpublished results). It is proposed that regular and reproducible multiple charging of the protein in vacuum in the context of the ESI method is possible because it retains its intact 3-dimensional fold (48). Taking this into account, it is possible to distinguish the five protein species in panel *B*: wt (35811.0 Da) as well as species with 1–4 chloro atoms (35847.2–35951.0 Da). Note also that the intensity of signal decreases with the increase of number of the chlorine atoms introduced, gradually leading to the decay of the structure. Standard deviation of the measurements did not exceed ±5.5 Da. Such behavior is reproducible for other proteins as well, for example, recombinant azurin (for experimental details, see Minks et al., unpublished results).

considerations can be found in a recent review of Richards (22), where it was suggested that the ''molecular packing, the efficient filling of the space, may be the most generally applicable factor that leads to the unique structures of most globular proteins.''

IN VIVO EXPERIMENTAL CODON REASSIGNMENT UNDER EFFICIENT SELECTIVE PRESSURE

Noncanonical amino acids that escape the natural editing mechanisms during translation are good candidates for efficient bioincorporation, since their integration will most probably result in sufficiently stable proteins. Protein building with an expanded amino acid repertoire relies on *in vivo* canonical amino acid codon reassignment to amino acids outside the canonical 20. Such assignment leads to residue-specific or codon-dependent replacements. In the laboratory, this can be achieved by incorporation under simple and efficient selective pressure in combination with a highly sophisticated and controllable protein expression technology. A particular auxotrophic host is exploited that is capable of efficient exclusion of possible toxic metabolic and physiological effects (16, 17, 21, 34–38). The commonly used procedure (selective pressure incorporation) is as follows: the metabolic pathway that supplies the cell with the particular canonical amino acid is switched off. The defined minimal media contain the noncanonical counterpart that is taken up by the cell into the cytoplasm, and there incorporation is forced despite possible physiological toxicity by a strong expression system that uses the whole cellular machinery to mainly express the target protein (37). Some examples are given in Table 1.

Although no systematic work in this direction has been performed, there are examples that confirm the simplicity of this *in vivo* approach. Experiments have been described aimed at generating suitable auxotrophic host cells, which are not only tolerant of canonical amino acids, but have even demonstrated better growth as shown for different fluorinated amino acids (38, 39). Phenotypic suppression that causes mischarging with the antibiotic amino acid azaleucine (whose structure mimics leucine) was also described (40). It was proposed that ''microbial strains with a clear-cut requirement for an additional amino acid should be instrumental for widening the genetic code experimentally.''

Recently, a high level of lysine for arginine misincorporations in recombinant eukaryotic gene expression due to the codon usage was demonstrated (41). In fact, codon usage as a measure of relative amounts of tRNA isoacceptors in two species often varies greatly for certain amino acids. For example, the arginine codon AGA corresponds to tRNA accounting for on average 50% of the total $tRNA^{Arg}$ available in the yeast cell, whereas this codon is the rarest among other arginine codons in *E. coli*, where it accounts for less than 4% of the total cellular $tRNA^{Arg}$ population (42). Such constrained codon usage can also act as efficient selective pressure. Indeed, it was demonstrated that, by proper media manipulations and the use of a strong expression system, it is possible to achieve a high level of basic amino acid lysine replacement with the similar amino acid arginine (42).

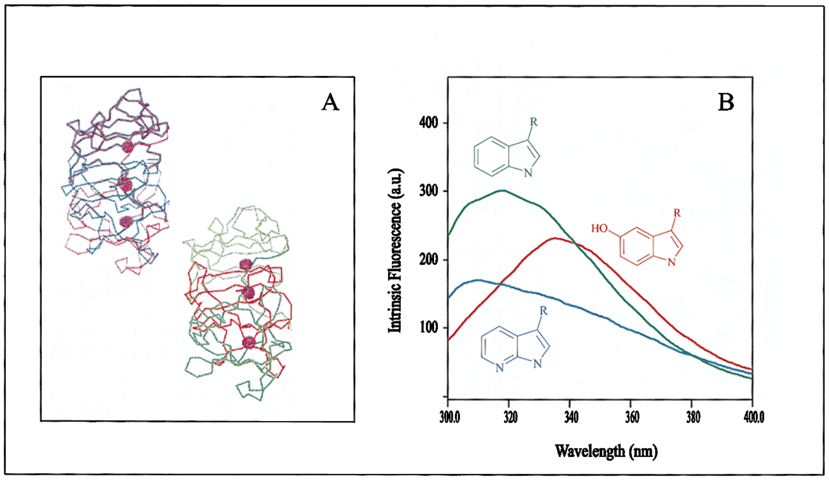

Figure 5. Examples for applications of noncanonical amino acid introduction into proteins. *A*) C^{α} representation of telluromethionine containing amino-terminal domain of tailspike protein shown by difference Fourier map and contoured at $6\,\sigma$. There are two trimmers (one methionine residue per monomer) in the asymmetric unit of the unit cell, where the difference density map indicates successful replacement of the canonical methionine with the noncanonical telluromethionine. The isomorphous presence of the metallic tellurium atom is heavy enough and scatters X-rays to a level that can be especially useful for the phase determination of the protein structure by protein X-ray crystallography (see ref 37 for more details). *B*) Fluorescence emission spectra of native (green), 5-hydroxy-tryptophan-containing (red), and 7-aza-tryptophan-containing (blue) human recombinant annexin V. All protein samples (2 μM) were excited at 280 nm in phosphate-buffered saline. Noncanonical amino acid-containing proteins have special spectral characteristics that distinguish them from the native ones. In general, proteins with 5-hydroxy-tryptophan exhibit a red shift of about 13 nm and a shoulder at 310 nm in the absorption spectrum. 7-Aza-tryptophan-containing proteins have an emission shoulder between 350 and 360 nm and reveal tyrosine contributions to the protein fluorescence emission spectrum at about 307 nm. For more details, see ref 36.

In all these cases, codon reassignment occurs by obeying a simple 'similar replaces similar' rule. In fact, an amino acid replacement model for code evolution, where chemically similar canonical amino acids are often coded by similar codons (1), further supports this prospect. For example, similar triplets (AUA, AUG) encode the canonical amino acids methionine and isoleucine, and indeed these two amino acids are comparably hydrophobic and replace one another in the phylogenic comparisons (43). It is easy to imagine that reassignment of one isoleucine codon to methionine might have occurred during evolution under certain conditions of strong selective pressure. In this way, the disruption of protein structure caused by codon meaning changes (codon capture) is minimized by such modest alterations in the amino acid side chain character. At the same time, the thioether moiety introduced into proteins allows for their further functional diversification (44).

PRACTICAL APPLICATIONS

By the 1950's, an interest was shown in the incorporation of noncanonical amino acids into proteins (45, 46). These were expected to serve as useful tools for studying mechanisms of protein synthesis and metabolism. However, only with the advent of recombinant DNA technology it was possible to take significant steps in the direction of an expanded amino acid repertoire. Therefore, this field is flourishing again, and practical implications of different approaches are invaluable and are extensively discussed elsewhere (5, 24). Two examples of some possibly useful practical applications for bioincorporation of noncanonical amino acids into proteins are given in **Fig. 5*A*, *B***.

POSSIBLE LESSONS FOR PROTEIN FOLDING: ATOMIC MUTATIONS

The current wide use of site-directed mutagenesis techniques for protein folding studies, although providing many useful insights, is limited because in that it is often difficult to unambiguously interpret the observed changes, since several sets of interactions are usually affected. The basic philosophy of our alternative proposal is very simple: the subtle replacement of 'natural' by 'nonnatural' should shed more light on the structural problem studied. Initial studies based on the replacement of Met by its noncanonical isosteres norleucine, selenomethionine, and telluromethionine resulted in a series of exchanges: $-CH_2- \rightarrow -S- \rightarrow -Se- \rightarrow -Te-$, i.e., in atomic mutations. These alterations allowed us to study the folding process as well as correlations between folding and physical-chemical characteristics of such atomic mutations (19). In addition, although such subtle exchanges result in crystallographically isomorphous protein structures (37), their thermodynamic behavior is different (16, 19). These results indicate that this new approach may represent a fine tool with which to study nonbonded interactions that are far below the resolution of conventional X-ray crystallographic and nuclear magnetic resonance protein analyses. Moreover, a hidden world of such interactions might easily emerge, revealing relationships responsible for the highly specific internal architecture and, thus, for the existence of a unique equilibrium structure of proteins.

THE BEST OF ALL POSSIBLE CODES?

Our living world is based on only one universal code, which is very conserved, rigid, stubborn, restricted, and resistant even to subtle changes. At the same time, de novo code engineering based on a new set of chemicals seems to be sheer fantasy. Thus, this code is probably the best of all codes accessible to our manipulations. The evolution based on it brought us here to continue and possibly speed up this process. Hard work is in front of us.

Sarah Teter, Richard Engh, Saulius Grazulis, and Karl-Peter Hopfner are greatly acknowledged for many invaluable suggestions, discussions and critical reading of this manuscript.

REFERENCES

1. Crick, F. H. C. (1968) The origin of the genetic code. J. Mol. Biol. 38, 367–369
2. Gellman, S. H. (1998) Foldamers: a manifesto. Acc. Chem. Res. 31, 173–180
3. Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., and Zinoni, F. (1991) Selenocysteine: the 21st amino acid. Mol. Microbiol. 5, 515–520
4. Jack, R. W., and Sahl, H. G. (1995) Unique peptide modifications involved in the biosynthesis of lantibiotics. TIBTECH 13, 269–278
5. Mendel, D., Cornish, V. W., and Schultz, P. G. (1995) Site-directed mutagenesis with an expanded genetic code. Annu. Rev. Biophys. Biomol. Struct. 24, 435–462
6. Cornish, V. W., Benson, D. R., Altenbach, C. A., Hideg, K., Hubbell, W. L., and Schultz, P. G . (1994) Site-specific incorporation of biophysical probes into proteins. Proc. Natl. Acad. Sci. USA 91, 2910–2914
7. Schimmel, P. (1996) Origin of genetic code: A needle in the haystack of tRNA sequences. Proc. Natl. Acad. Sci. USA 93, 4521–4522
8. Edelman, P., and Callant, J. (1977) Mistranslation in *E. coli.* Cell 10, 131–137
9. Brick, P., and Blow, D. M. (1987) Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine. J. Mol. Biol. 194, 287–297
10. Fersht, A. R. (1981) Enzymatic editing mechanisms and the genetic code. Proc. Roy. Soc. London B: Biol. Sci. 212, 351–379
11. Fersht, A. R., and Dingwall, C. (1979) An editing mechanism for the methionyl-tRNA synthetase in the selection of amino acids in protein synthesis. Biochemistry 18, 1250–1256
12. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. Science 181, 223–230
13. Rose, G. D., and Wolfenden, R. (1993) Hydrogen bonding, hydrophobicity, packing and protein folding. Annu. Rev. Biophys. Biomol. Struct. 22, 381–415
14. Kudlicki, W., Chirgwin, J., Kramer, G., and Hardesty, B. (1995) Folding of an enzyme into an active conformation while bound as peptidyl-tRNA to the ribosome. Biochemistry 34, 14284–14287
15. De Marco, C., Busiello, M., Di Girolamo, M., and Cavallini, D. (1977) Selenaproline and protein synthesis. Biochim. Biophys. Acta 478, 156–166
16. Budisa, N., Minks, C., Medrano, F. J., Lutz, J., Huber, R., and Moroder, L. (1998) Residue specific bioincorporation of non-natural, biologically active amino acids into proteins as possible drug carriers: structure and stability of the per-thiaproline mutant of annexin V. Proc. Natl. Acad. Sci. USA 95, 455–459
17. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J., and Huber, R. (1995) High level biosynthetic substitution of methionine in proteins by its analogues 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in *Escherichia coli.* Eur. J. Biochem. 230, 788–796
18. Osawa S., Jukes T. H., Watanabe, K., and Muto, A. (1992) Recent evidence for evolution of the genetic code. Microbiol. Rev. 56, 229–264
19. Budisa, N., Huber, R., Golbik, R., Minks, C., Weyher, E., and Moroder, L (1998) Atomic mutations in annexin V. Thermodynamic study of isomorphous protein variants. Eur. J. Biochem. 253, 1–9
20. Wong, J. T. F. (1998) Evolution of the genetic code. Microbiol. Sci. 5, 174–181
21. Beiboer, S. H. W., Van den Berg, B., Dekker, N., Cox, R. C., and Verhej, H. M. (1996) Incorporation of an unnatural amino acid in the active site of porcine pancreatic phospholipase A2. Substitution of histidine by 1,2,4-triazole-3-alanine yields an enzyme with high activity at acidic pH. Protein Eng. 9, 345–352
22. Richards, F. M. (1997) Protein stability: still an unsolved problem. CMLS 53, 790–802
23. Rogers, K. C., and Söll, D. (1995) Divergence of glutamate and glutamine aminoacylation pathways: Providing the evolutionary rationale for mischarging. J. Mol. Evol. 40, 476–481
24. Ibba, M., and Hennecke, H. (1994) Towards engineering proteins by site-directed incorporation *in vivo* of non-natural amino acids. Bio/Technology 12, 678–682
25. Baron, C., and Böck, A. (1995) The selenocysteine inserting tRNA species: structure and function. In: tRNA: Structure, Biosynthesis and Function (Söll, D., and RajBhandary, U., eds) pp. 529–544, American Society for Microbiology, Washington, D.C.,
26. Huber, R. E., and Criddle, R. S. (1967) Comparison of the chemical properties of selenocysteine and selenocystine with their sulfur analogs. Arch. Biochem. Biophys. 122, 164–173
27. Besse, D., Budisa, N., Karnbrock, W., Minks, C., Musiol, H. J., Pegoraro, S., Siedler, F., Weyher, E., and Moroder, L. (1997) Chalcogen-analogs of amino acids. Their use in X-ray crystallographic and folding studies of peptides and proteins. J. Biol. Chem. 378, 211–218
28. Ladenstein, R., Epp, O., Bartels, K., Jones, A., Huber, R., and Wendel, A. (1979) Structure analysis and molecular model of the selenoenzyme glutathione peroxidase at 2.8 Å resolution. J. Mol. Biol. 134, 199–218
29. Schmidt, E., and Schimmel, P. (1993) Dominant lethality by expression of a catalytically inactive class I tRNA synthetase. Proc. Natl. Acad. Sci. USA 90, 6919–6923
30. Schimmel, P., and Söll, D. (1997) When protein engineering confronts the tRNA world. Proc. Natl. Acad. Sci. USA 94, 10007–10009
31. Liu, D. R., Magliery, T. J., Pasternak, M., and Schultz P. G. (1997) Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins *in vivo.* Proc. Natl. Acad. Sci. USA 94, 10092–10097
32. Ibba, M., and Hennecke, H. (1995) Relaxing the substrate specificity of an aminoacyl-tRNA synthetase allows *in vitro* and *in vivo* synthesis of proteins containing unnatural amino acids. FEBS Lett. 364, 272–275
33. Craig, A. G., Jimenez, E. C., Dykert, J., Nielsen, D. B., Gulyas, J., Abogadie, F. C., Porter, J., Rivier, J. E., Cruz, L. J., Olivera, B. M., and McIntosh, J. M. (1997) A novel post-translational modification involving bromination of tryptophan. Identification of the residue, L-6-bromotryptophan, in peptides from Conus imperialis and Conus radiatus venom. J. Biol. Chem. 272, 4689–4698
34. Yoshikawa, E., Fournier, M. J., Mason, T. L., and Tirrell, D. A. (1994) Genetically engineered fluoropolymers. Synthesis of repetitive polypeptides containing *p*-fluorophenylalanine residues. Macromolecules 27, 5471–5475
35. Kothakota, S., Mason, T., Tirrell, D. A., and Fournier, M. J. (1995) Biosynthesis of a periodic protein containing 3-thienylalanine: a step toward genetically engineered conducting polymers. J. Am. Chem. Soc. 117, 536–537
36. Ross, J. B., Szabo, A. G., Hogue, C. W. (1997) Enhancement of protein spectra with tryptophan analogs: fluorescence spectroscopy of protein–protein and protein–nucleic acid interactions. Methods Enzymol. 278, 151–190
37. Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neuefeind, T., Moroder, L., and Huber, R. (1997) Bioincorporation of telluromethionine into proteins: a promising new approach for X-ray structure analysis of proteins. J. Mol. Biol. 270, 616–623
38. Ring, M., Armitage, I. M., and Huber, R. E. (1985) *m*-Fluorotyrosine substitution in β-galactosidase: evidence for the existence of a catalytically active tyrosine. Biochem. Biophys. Res. Commun. 131, 675–680

39. Bronskill, P. M., and Wong, J. T. (1988) Suppression of fluorescence of tryptophan residues in proteins by replacement with 4-fluorotryptophan. Biochem. J. 249, 305–308

40. Lemeignan, B., Sonigo P., and Marliere, P. (1993) Phenotypic suppression by incorporation of an alien amino acid. J. Mol. Biol. 231, 161–166

41. Calderone, T. L., Stevens, R. D., and Oas, T. G. (1996) High level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in *Escherichia coli.* J. Mol. Biol. 262, 407–412

42. Forman, M. D., Stack, R. F., Masters, P. S., Hauer, C. R., and Baxter, S. M. (1998) High level, context dependent misincorporation of lysine for arginine in *Saccharomyces cerevisiae* a1 homeodomain expressed in *Escherichia coli.* Protein Sci. 7, 500–503

43. Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, Vol. 5. National Biomedical Research Foundation, Washington D.C.

44. Gelman, S. (1991) On the role of methionine residues in the sequence-independent recognition of non-polar protein surfaces. Biochemistry 30, 6633–6636

45. Cowie, D. B., and Cohen, G. N. (1957) Biosynthesis by *Escherichia coli* of active altered proteins containing selenium instead sulphur. Biochim. Biophys. Acta 26, 252–261

46. Richmond, M. H. (1962) The effect of amino acid analogues on growth and protein synthesis in microorganisms. Bacteriol. Rev. 26, 398–420

47. Tsugita, A., Uchida, T., Mewes, H. W., and Ataka, T. (1987). A rapid vapour phase acid (hydrochloric and trifluoroacetic acid) hydrolysis of peptide and protein. J. Biochem. 102, 1593–1597

48. Mann, M., and Wilm, M. (1995) Electrospray mass spectrometry for protein characterisation. Trends Biochem. Sci. 20, 219–223

49. Müller, S., Senn, H., Gsell, B., Vetter, W., Baron, C., and Böck, A. (1994) The formation of diselenide bridges in proteins by incorporation of selenocysteine residues: biosynthesis and characterization of (Se)2-thioredoxin. Biochemistry 33, 3404–3412

50. Wong, C. Y., and Eftink, M. R. (1998) Incorporation of tryptophan analogues into staphylococcal nuclease, its V66W mutant, and d137–149 fragment: spectroscopic studies. Biochemistry 37, 8938–8946

Received for publication August 13, 1998.
Revised for publication September 8, 1998.