Abstract: We have generated a completely autonomous bacterium with a 21 amino acid genetic code. This bacterium can biosynthesize a nonstandard amino acid from basic carbon sources and incorporate this amino acid into proteins in response to the amber nonsense codon. The biosynthetic pathway for the amino acid \( p \)-aminophenylalanine (\( p \)AF) as well as a unique \( p \)AF synthetase and cognate tRNA were added to \textit{Escherichia coli}. Denaturing gel electrophoresis and mass spectrometric analysis show that \( p \)AF is incorporated into myoglobin with fidelity and efficiency rivaling those of the common 20 amino acids. This and other such organisms may provide an opportunity to examine the evolutionary consequences of adding new amino acids to the genetic repertoire, as well as generate proteins with new or enhanced biological functions.

Introduction

The genetic code, consisting of four nucleotides specifying 64 unique triplet codons, encodes the 20 amino acid building blocks from which all proteins are synthesized.\(^1\) Is this the ideal number of basic building blocks or would additional amino acids lead to proteins or even organisms with enhanced function? Although various arguments have been put forth to explain the nature and number of amino acids in the code,\(^2\)\(^-\)\(^4\) it is clear that proteins require additional factors to carry out many of their natural functions.\(^5\) These groups are provided through post-translational modification and cofactors,\(^6\)\(^,\)\(^7\) and in rare instances, organisms have evolved novel translational machinery to incorporate either selenocysteine or pyrrolysine in response to nonsense codons.\(^8\)\(^-\)\(^10\) The need for additional chemistries in so many protein-mediated processes raises the question of why the genetic code has not evolved further.

The generation of bacteria with expanded genetic codes containing additional amino acid building blocks would provide an experimental system to address this central question of evolution. To date there have been a number of examples in which one of the common 20 amino acids has been substituted to varying extents with close analogues, including fluorine, nitrogen, and selenium substitution.\(^11\)\(^-\)\(^15\) In addition, \textit{in vitro} biosynthetic methods using chemically aminocacylated tRNAs and peptide synthesis have allowed additional amino acids to be incorporated into proteins to study protein structure and function. However, none of these methods have made it possible to genetically encode additional amino acids beyond the common 20 in living organisms.

Recently, we showed that one could add new components to the translational machinery of \textit{Escherichia coli} to incorporate additional amino acids added exogenously to the growth media into proteins with high selectivity and fidelity.\(^16\)\(^-\)\(^18\) We now report the construction of a completely autonomous bacterium that genetically encodes the additional amino acid \( p \)-aminophenylalanine (\( p \)AF). This bacterium has (i) the ability to synthesize \( p \)AF from simple carbon sources, (ii) an aminocaclyt-tRNA synthetase that uniquely utilizes \( p \)AF and no other amino acid, and (iii) a tRNA that is acylated by this synthetase and no other, and that delivers \( p \)AF efficiently into proteins in response to the amber codon, TAG. The amino acid is incorporated with fidelity rivaling the common 20 amino acids.

...
Results and Discussion

Selection of a pAF Synthetase. For the purposes of these experiments, we chose the nonproteinogenic amino acid p-aminophenylalanine (1, pAF), on the basis of its interesting physical properties (e.g., π-donating effects, hydrogen-bonding properties, and weak basicity), stability, and lack of toxicity to E. coli, and the fact that it is a known secondary metabolite. Moreover, the genes that lead to the production of pAF as a metabolic intermediate in the production of chloramphenicol and pristinamycin have been identified in Streptomyces venezuelae and Streptomyces pristinaeaspiralis, respectively.19,20

To generate an orthogonal tRNA-synthetase pair for pAF, the Methanococcus jannaschii tyrosyl-tRNA synthetase (TyrRS) and mutant tyrosine amber suppressor tRNA (mutRNA CUA Tyr) pair was used as a starting point.21,22 Natural and synthetic amber codon suppressors have been extensively used to mutate proteins with little negative effect on cell growth.23,24 The pAF-specific synthetase (pAFRS) was generated by modifying the amino acid specificity of the M. jannaschii TyrRS to accept pAF and not any of the common 20 amino acids. A combination of positive and negative selections25 was used to identify the pAFRS enzyme from a library of TyrRS variants containing random amino acids at five positions (Tyr32, Glu107, Asp158, Ile159, and Leu162).16

A single reporter plasmid was used for both selection and screening. The positive selection is based on suppression of a TAG codon at a permissive position within the chloramphenicol acetyltransferase (CAT) gene16,26 by either pAF or an endogenous amino acid. Cells containing the TyrRS library and reporter plasmid were grown in liquid culture containing pAF and selected for survival in the presence of chloramphenicol (Cm). The negative screen is based on the inability to suppress in the absence of pAF two TAG stop codons at permissive positions within the T7 RNA polymerase gene. Production of full-length T7 RNA polymerase drives production of a variant of green fluorescent protein (GFPuv). Cells from the positive selection were grown in the absence of pAF and Cm, and then screened using fluorescence-activated cell sorting for a lack of fluorescence. The desired pAFRS was identified following two rounds of positive selection in liquid media, one round of negative screening, another round of positive selection in liquid media, and one round of positive selection on plates. The pAFRS enzyme contains five mutations relative to the wild-type TyrRS (Y32T, E107T, D158P, I159L, and L162A). In the absence of pAF, the IC50 of cells containing the selected pAFRS and reporter plasmid was 10 μg/mL. Cm. The IC50 was 120 μg/mL Cm with 1 mM pAF. This result suggests that pAF is necessary for suppression of the TAG codon.

Engineering Nonproteinogenic Amino Acid Production. We anticipated that pAF could be synthesized in E. coli from chorismate 2 (a biosynthetic intermediate in the synthesis of aromatic amino acids) using the S. venezuelae enzymes PapA, PapB, and PapC, together with an E. coli aminotransferase (Figure 1B). PapA, 4-amino-4-deoxypyruvate synthase, is expected to convert 2 to 4-amino-4-deoxypyruvate (3) using ammonia (from glutamine) in a simple addition–elimination reaction. PapB and PapC, which are analogous to chorismate mutase and prephenate dehydrogenase, respectively, should convert 3 to 4-amino-4-deoxychorismic acid (4) and then to 5. The non-specific tyrosine aminotransferases from E. coli (tyrB, aspS, and ileE) have been shown to utilize many ketoacids and were expected to convert 5 to pAF.27

It was difficult to predict how the synthesis of another aromatic amino acid from the heavily used chorismate branch point would affect E. coli growth. p-Aminophenylalanine must be produced in a concentration sufficient for efficient protein biosynthesis, but not to such a degree as to affect the concentration of the other aromatic amino acids or to exhaust cellular resources. In S. venezuelae there is evidence to suggest that the regulation of the shikimate pathway is modified to account for chorismate consumption in making a fourth aromatic amino acid.28 The availability of an orthogonal pAFRS/mutRNA CUA Tyr

Figure 1. S. venezuelae genes papA, papB, and papC cloned into pSC101-derived pLASC plasmid under control of the lac or lpp promotor (A). Proteins PapA, PapB, and PapC convert chorismate 2 to p-aminophenylpyruvate (5), while E. coli aromatic aminotransferase completes the biosynthesis to afford p-AF (1) (B).
pair that can selectively incorporate pAF into proteins should allow one to optimize the in vivo production of pAF. Only cells that produce a concentration of pAF sufficient for protein biosynthesis will be able to suppress the amber codon in the CAT gene and survive on Cm, or suppress the two TAG codons in the T7 RNA polymerase gene and produce GFPuv. One can use these suppression assays to monitor pAF biosynthesis under different E. coli growth conditions and select for optimal pAF production on the basis of E. coli growth rates.

To incorporate the pAF biosynthetic pathway into E. coli, the papABC genes from S. venezuelae were cloned into a low-copy pSC101-derived plasmid (Figure 1A). The gene cassettes were placed under the control of the constitutive lpp promoter, or the inducible lac promoter, and each gene was placed behind an E. coli ribosome binding site (rbs). The ability of the cells to generate pAF and suppress a TAG stop codon was first monitored by observing fluorescence due to the production of GFPuv. Both the lpp promoter-controlled pAF biosynthetic pathway (lppPW) and the addition of 1 mM exogenous pAF produced 120-fold higher cellular fluorescence than the background (fluorescence when no pAF is available to the GFPuv reporter system), while the lac promoter-controlled pAF biosynthetic pathway (lacPW) without IPTG induction produced a 3.2-fold increase in cellular fluorescence. The GFPuv suppression assay shows that biosynthesis of pAF by the lppPW is comparable to that by exogenous 1 mM pAF, and significantly better than that by the noninduced lacPW.

To determine whether the addition of the pAF biosynthetic pathway affects the production of other aromatic amino acids in E. coli, and to quantify pAF production, the cellular concentrations of the aromatic amino acids were monitored. Cells with lppPW or lacPW were compared to those without the biosynthetic pathway, grown either in the presence or absence of exogenous pAF. The cellular amino acid concentrations were determined by extracting cells at midlog growth followed by HPLC–MS analysis. The concentrations of the natural aromatic amino acids remained unchanged under all conditions examined, indicating that neither the presence of the pAF pathway nor exogenous pAF significantly affects amino acid production (Table 1). While both pathways produce pAF in concentrations similar to those of the natural aromatic amino acids, lppPW produces 7 times as much pAF (0.7 mM) as the noninduced lacPW (0.1 mM).

Production of pAF-Containing Protein by an Engineered Bacterium. To confirm that the pAF biosynthetic pathway and mutRNA_CUA\(^{\text{Trp}}\)/pAFRS pair lead to incorporation of pAF (and only pAF) into proteins only in response to TAG codons, protein containing pAF was expressed, purified, and analyzed. The fourth codon of the sperm whale myoglobin gene was converted to a TAG stop codon, and a COOH-terminal His6-tag was added to facilitate purification. When pAFRS was coexpressed with the mutRNA_CUA\(^{\text{Trp}}\) in the presence of 1 mM pAF, or the genes coding for the amino acid biosynthetic pathway (lppPW or lacPW), full-length myoglobin was produced (Figure 2). No myoglobin was observed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) and silver staining in the absence of pAFRS or a pAF source, indicating that only pAF is inserted at the fourth position of myoglobin. Coproduction of the mutant synthetase with lppPW, lacPW, or exogenously added pAF yielded 1.8, 0.2, and 3.0 mg/L mutant protein in liquid GMML, respectively. To compare the efficiency of pAFRS and the biosynthetic pathway, the wild-type TyrRS was used to make 4-Tyr-myoglobin, and 2.2 mg/L protein was produced. Wild-type myoglobin was produced and purified under the same conditions as 4-pAF-myoglobin to determine the suppression efficiency of the pAFRS. The yield of wild-type myoglobin was 3.5 mg/L, comparable to the 3.0 mg/L yield when pAFRS was used to suppress the amber codon in the fourth position of myoglobin. This indicates that amber suppression with the pAFRS is not the rate-limiting step in protein production.

To further demonstrate that only pAF is inserted in response to the TAG codon, the purified myoglobin protein was analyzed by N-terminal sequencing. Edman degradation showed that the fourth position of 4-pAF-myoglobin was occupied by only a nonstandard amino acid and none of the 20 natural amino acids within the limits of detection (10–20%). To verify that the fourth amino acid of the modified myoglobin is pAF, the protein mass was analyzed by electrospray-ionization ion trap mass spectrometry. The average protein mass resulting from exogenously added pAF or biosynthetically synthesized pAF was 18430.01 (SD 0.56) or 18430.00 (SD 0.45) Da, respectively (calculated mass 18430.2 Da). If the wild-type TyrRS is used with the mutRNA_CUA\(^{\text{Trp}}\), then tyrosine is inserted instead of pAF, and a mass shift of 1 Da higher is observed. To further assess whether this mass shift of 1 Da was due to the discrete incorporation of biosynthesized pAF into the fourth position of myoglobin, isolated proteins were digested with trypsin and the fragments were analyzed by MALDI-FT-ICR (MS). The sequence coverages of 4-pAF-myoglobin and 4-Tyr-myoglobin were greater than 85% and 79%, respectively, with a 1 Da shift detectable only in the first 17 amino acid peptide fragment (MVL(pAF)EGWLVHLVWAK). The monoisotopic singly protonated peptide mass for the 4-pAF-containing protein was 2100.094 ppm (theoretical M + H, 2100.115 ppm) with a 2 ppm error in measured mass, while that of the 4-Tyr-containing protein was 2088.138 ppm (theoretical M + H, 2088.115 ppm) with a 2 ppm error in measured mass.

**Table 1.** Cellular Concentrations (mM) of Aromatic Amino Acids in E. coli.\(^{\text{a}}\)

<table>
<thead>
<tr>
<th>amino acid</th>
<th>pLASC</th>
<th>pLASC + exogenous pAF</th>
<th>pLASC-lacPW</th>
<th>pLASC-lppPW</th>
</tr>
</thead>
<tbody>
<tr>
<td>pAF</td>
<td>0.0</td>
<td>2.5</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.8</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Trp</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Phe</td>
<td>0.3</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

\(^{\text{a}}\) Cells without the pAF biosynthetic pathway (pLASC), cells with the pathway under the control of the lac promoter (pLASC-lacPW), and cells with the pathway under the control of the strong lpp promoter (pLASC-lppPW) were grown in minimal media. Amino acids were extracted by toluenization, separated by HPLC, and detected by mass spectrometry.
peptide fragment was 2101.096 ppm (theoretical M + H, 2101.094 ppm) with a 1 ppm error in the measured mass. These data, together with the SDS–PAGE analysis, conclusively show that pAF is being incorporated with high fidelity and efficiency.

**Conclusion**

We have generated for the first time a completely autonomous organism that genetically encodes a novel amino acid that can be incorporated into proteins efficiently and with high fidelity in a manner analogous to that of the 20 standard amino acids. Although the additional genes are encoded on low copy plasmids, it is likely that they could be incorporated directly into the *E. coli* genome. Moreover, there are no observable toxic side effects due to amber codon suppression, as expected from previous studies.23,24 Efforts are in progress to generate a completely unique codon for pAF by using four-base codons to encode additional amino acids. However, with the current version of the pAF bacterium, we can now begin to compare its ability to adapt to selective pressure to that of its 20 amino acid ancestor. Moreover, by evolving this bacterium and others like it in response to selective pressures, we hope to determine whether the addition of novel amino acids to the genetic code can provide an evolutionary advantage to *E. coli*.

**Experimental Section**

**Evolution of a Mutant Synthetase Specific for p-Aminophenylalanine (pAFRS).** The pAF amino acid was purchased from Sigma. The starting library consists of *M. jannaschii* tyrosyl-tRNA synthetase (TyrRS) variants, randomized at positions Tyr2, Gly107, Asp158, Ile159, and Leu162,16 and inserted into plasmid pBK-JYRS, which contains the ColE1 origin of replication and a kanamycin (Kn) resistance marker. The reporter plasmid pREP/YC-JYCUA25 contains the genes for tyrosine amber suppressor tRNA (mutRNACUA Tyr ), and a selectable marker for tetracycline (Tet) resistance. The CAT gene contains an overlap PCR using the three cloned genes and primers 5′-CCGTT-TGAATTCCACACAGGAAACAGC-3′ forward, 5′-ACAACCTCTTGTAGATCAGGCGGCTATCCACAG-GTCCTCG-3′ reverse (papB); 5′-CACACAAAGGACCATGACATGACGAGGTCGCT-3′ forward, 5′-ACAACCTCTTGTAGATCAGGCGGCTATCCACAG-GTCCTCG-3′ reverse (papC). Genes *papABC* were assembled by overlap PCR using the three cloned genes and primers 5′-CCGTT-TGAATTCCACACAGGAAACAGC-3′ forward and 5′-GGAAGGATCCCTATAGTGTCACTAAAT-3′ reverse. The papABC fragment digested with BamHI and EcoRI was inserted into a pSC101-derived plasmid, pLASC, and maintained by ampicillin (Amp) selection. Ribosome binding sites were derived from the 5′-UTR of LacZ (5′-CACACAGGAAACAGC-3′), malE (5′-CACACAAAGGACCATGAC-3′), and cro (5′-AGTACTAAGGCGTGT-3′) and placed prior to *papA*, *papB*, and *papC*, respectively. The *papABC* genes were placed under control of lac and lpp promoters to afford two pathway plasmids: pLASC-lacPW and pLASC-lppPW (as shown Figure 1A).

**Testing pAF Biosynthesis with pAFRS.** *E. coli* DH10B cells harboring three plasmids, the reporter plasmid pREP/YC-JYCUA, the synthetase plasmid pAFRS, and the pathway plasmid pLASC-lacPW or pLASC-lppPW, were grown to saturation in GMML containing 100 μg/mL Amp (the pLASC plasmid was used in place of the pathway plasmid for the background, no pAF, and 1 mM exogenous pAF trials). DH10B was grown with no plasmids to determine the background suppression level of the reporter plasmid. A sample of each cell growth was equalized to an OD of 1.0 (600 nm) with water, washed, and diluted 5-fold with PBS. Cells suspended in PBS were analyzed using a Fluoromax-2 fluorescence detector (the excitation wavelength was 351 nm, and a peak emission at 505 nm was monitored). DH10B produced 1.0 × 10^6 fluorescence units, while background fluorescence (no pAF added) from the reporter system produced 2.5 × 10^6 fluorescence units. The lacPW, lppPW, and 1 mM exogenously added pAF produced 7.9 × 10^6, 3.0 × 10^6, and 3.0 × 10^6 fluorescence units, respectively. Induction of the lacPW with IPTG was not feasible due to its inhibitory affect on the arabinose promoter in the reporter plasmid pREP/YC-JYCUA.

**Aromatic Amino Acid Concentration.** *E. coli* DH10B cells harboring the pLASC plasmid, pLASC-lacPW, or pLASC-lppPW were grown to saturation in GMML minimal medium containing 110 μg/mL Amp. Cells grown with exogenously added pAF contained 1 mM amino acid at the start of the growth. Cells (100 mL, grown to saturation) were harvested by centrifugation and washed (1 mL of water), and 1 mL of water and 0.2 mL of toluene were added. The cell pellet was suspended in the extraction mixture, shaken at 37 °C for 30 min, and then separated by centrifugation. The aqueous layer was filtered (Microcon YM-10) and analyzed by HPLC. Cells grown with 0.1 mM, while for pAF and tryptophan, and phenylalanine to the extraction mixture and effectively back-calculate the original concentration of the amino acids present. The error in the amino acid concentration for pAF and tryptophan is ±0.1 mM, while for

D J. AM. CHEM. SOC.
tyrosine and phenylalanine it is ±0.2 mM. Cellular concentrations were based on the amount of water in the cell pellet (assuming the hydrated pellet was 70% water by mass).

**Production of Protein Containing pAF.** The plasmid pBAD/JYAMB-4TAG (Tet resistance) was used to express the mutRNA_{CUA}^{Tyr} gene under the control of the lpp promoter and rrnC terminator, and the sperm whale myoglobin gene (with an amber stop codon at Ser4) under the control of the arabinose promoter and rrnB terminator. The pBAD/JYAMB plasmid was used to express the wild-type myoglobin gene (with no amber stop codon) to compare protein yields to determine suppression efficiency. A his6-tag was added to the carboxy terminus of myoglobin. The TyrRS and pAFRS genes were expressed under the control of the *E. coli* GlnRS promoter and terminator on a pBR322-derived plasmid with Kn resistance. The *papABC* genes were expressed from pLASC-lacPW or pLASC-lppPW under the control of the native terminator. *E. coli* DH10B cells harboring the plasmid pBAD/JYAMB-4TAG, pBK-TyrRS, or pBK-pAFRS and a pLASC-derived plasmid (pLASC, pLASC-lacPW, or pLASC-lppPW as indicated) were grown in 0.5 L of GMML minimal media containing 5 μM FeCl₃. Protein production trials with exogenous pAF contained a final concentration of 1 mM pAF. For all trials, arabinose was added to a final concentration of 0.002% when the cells reached an OD of 0.3−0.4. The cultures were all grown to saturation (20−30 h) in parallel at 37 °C and pelleted, and protein was purified by Ni²⁺ affinity chromatography according to the manufacturer’s protocol under native conditions (Qiagen, Valencia, CA). A 15 μL sample of the final protein solution (3.5 mL) from each preparation was separated on a 12% SDS–polyacrylamide gel and quantified by silver staining.

**Characterization of Myoglobin Samples.** All myoglobin samples, with pAF in the fourth position (4-pAF-myoglobin), with tyrosine in the fourth position (4-Tyr-myoglobin), and wild-type myoglobin with serine in the fourth position, were purified prior to analysis. Protein samples for Edman sequencing, trypsin digests, and mass spectrometry were purified first by Ni²⁺ affinity chromatography and then by HPLC (Dynamax). Desalted myoglobin samples from affinity chromatography were separated on a C4 column (5 μm, 4.6 × 150 mm) with a gradient of aqueous 0.1% TFA/acetonitrile 0.1% TFA (80:20 to 10:90) over 30 min. Samples were lyophilized and used for analysis.

**Acknowledgment.** We acknowledge support of this work by the Department of Energy (Contract No. DE-FG03-00ER45812). J.C.A. is an NSF predoctoral fellow. S.W.S. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research. This is Manuscript No. 14864-CH of The Scripps Research Institute.

**Supporting Information Available:** MALDI-FT-ICR (MS) analysis of the N-terminal tryptic fragments of 4-Tyr-myoglobin and 4-pAF-myoglobin (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

JA0284153