Protein Science (1998), 7:500–503. Cambridge University Press. Printed in the USA.

FOR THE RECORD

High level, context dependent misincorporation of lysine for arginine in *Saccharomyces cerevisiae* a1 homeodomain expressed in *Escherichia coli*

MICHAEL D. FORMAN, ROBERT F. STACK, PAUL S. MASTERS,
CHARLES R. HAUER, AND SUSAN M. BAXTER
The Wadsworth Center, New York State Department of Health, Empire State Plaza, Albany, New York 12201-0509

(RECEIVED August 19, 1997; ACCEPTED November 23, 1997)

Abstract: The *Saccharomyces cerevisiae* **a**1 homeodomain is expressed as a soluble protein in *Escherichia coli* when cultured in minimal medium. Nuclear magnetic resonance (NMR) spectra of previously prepared **a**1 homeodomain samples contained a subset of doubled and broadened resonances. Mass spectroscopic and NMR analysis demonstrates that the heterogeneity is largely due to a lysine misincorporation at the arginine (Arg) 115 site. Arg 115 is coded by the 5′-AGA-3′ sequence, which is quite rare in *E. coli* genes. Lower level mistranslation at three other rare arginine codons also occurs. The percentage of lysine for arginine misincorporation in **a**1 homeodomain production is dependent on media composition. The *dnaY* gene, which encodes the rare 5′-AGA-3′ tRNA^{ARG}, was co-expressed in *E. coli* with the **a**1-encoding plasmid to produce a homogeneous recombinant **a**1 homeodomain. Co-expression of the *dnaY* gene completely blocks mistranslation of arginine to lysine during **a**1 overexpression in minimal media, and homogeneous protein is produced.

Keywords: AGA; low-use codons; NMR spectroscopy; mass spectroscopy; mistranslation; protein heterogeneity

For high resolution NMR structural studies, proteins from yeast and other eukaryotic organisms are commonly expressed in *Escherichia coli* to maximize expression and to isotopically label at reasonable cost. It is critically important to develop an expression system that is not only efficient, but also accurate. Although the expression systems of yeast and *E. coli* are similar, codon usage in the two species varies greatly for certain amino acids (Sharp et al., 1986; Kane, 1995). In yeast, the preferred codon for the arginine (Arg) residue is 5′-AGA-3′, with its corresponding tRNA accounting for over 50% of the total tRNA^{ARG} available in the cell. In *E. coli,* the preferred codon for the Arg residue is 5′-CGC-3′ or 5′-CGU-3′. The 5′-AGA-3′ codon is the rarest codon used by *E.coli*, in terms of tRNA availability, accounting for only 4% of the total tRNA^{ARG} population (Wada et al., 1992). Thus, expression of yeast proteins in *E. coli* may be constrained by codon usage. Cases of reduced levels of protein expression, plasmid instability, and partial mistranslation at these rare codons sites have been reported (Seetharam et al., 1988; Kane, 1995; Zahn, 1996). Recent reports (Calderone et al., 1996; Day et al., 1996) have demonstrated that levels of lysine for arginine misincorporation in recombinant eukaryotic gene expression can be much higher than previously thought, but the mechanisms underlying high-level misincorporation have not been elucidated.

The *Saccharomyces cerevisiae* **a**1 homeodomain (**a**1$_{66-126}$) has been produced for NMR studies in soluble form using an *E. coli* expression system. NMR spectra of recombinant **a**1 homeodomain contain a number of peaks that are broadened and even doubled (Baxter et al., 1994), suggesting that the protein produced is not homogenous. Heterogeneous protein, as judged from NMR spectra, is obtained from a series of preparations using a variety of *E. coli* strains, culture media, and growth conditions. Electrospray mass spectra (ESI-MS) of a series of **a**1 homeodomain samples show the major component at the expected mass (*M*), but also significant amounts of impurities with unexpected mass differentials of (M-28) and (M-56). Peptide digestion and mass spectroscopic sequence analysis determined the source of the impurities to be misincorporated lysine at arginine codons. Two alternative expression systems were developed to block mistranslation at the 115 codon, the major site of lysine misincorporation. The first approach involved replacing the rare AGA codon at site 115 with the arginine codon more commonly found in *E. coli*, CGC. The second system involved coexpression of the *dnaY* gene (Garcia et al.,

Reprint requests to: Susan M. Baxter, The Wadsworth Center, New York State Department of Health, Empire State Plaza, Albany, NY 12201-0509; e-mail: baxter@wadsworth.org.

Abbreviations: NMR, nuclear magnetic resonance; HSQC, heteronuclear single quantum correlation; ESI-MS, electrospray ionization mass spectroscopy; MALDI-TOF, matrix-assisted laser desorption ionization time-of-flight; LB, Luria-Bertini culture broth; **a**1 homeodomain, **a**1$_{66-126}$

1986; Brinkmann et al., 1989), which codes for the rare 5′-AGA-3′ tRNA^{ARG}, along with the original **a1** coding plasmid. ESI-MS analysis demonstrates that replacement of the rare codon at 115 results in nearly pure protein, but that low-level misincorporation is still occurring at alternate rare codons in the sequence. Coexpression of the *dnaY* gene completely blocks mistranslation as shown by ESI-MS and NMR analysis.

Results and discussion: Broadened and doubled peaks in NMR spectra of the **a1** homeodomain (**a1**$_{66-126}$) were observed in all preparations of uniformly ^{15}N-labeled recombinant protein (Baxter et al., 1994). The **a1** homeodomain is expressed in *E. coli*, cultured in minimal medium, in soluble form at 8 mg/L levels. An expansion of the ^{1}H-^{15}N HSQC spectrum of a typical uniformly ^{15}N-labeled **a1** homeodomain sample is shown in Figure 1A. Doubled peaks include R115, W117, and I119, which are contained in the C-terminal, DNA-binding helix of the homeodomain. Residues not included in the C-terminal helix but close in three-dimensional space, such as A104, are also doubled. Chromatographic and gel electrophoretic methods did not reveal the presence of multiple protein species in the preparations.

We used ESI-MS to identify the source of the peak doubling in an **a1** homeodomain preparation that contained approximately 15% impurity, as judged by integration of doubled crosspeaks in an [^{15}N-^{1}H] HSQC NMR spectrum. Using reversed phase high performance liquid chromatography-electrospray ionization mass spectroscopy, three components were identified in the mass spectrum of a representative sample, shown in Figure 2A. The major component was at the expected mass (M = 7,277 Da) for correctly expressed protein with an N-terminal methionine. Minor components, making up about 20% of the sample population, represented protein with a −28 Da mass differential (M-28) and −56 Da mass differential (M-56). Enzymatic digestion and mass spectroscopic analysis via ESI-MS/MS and MALDI-TOF revealed that the (M-28) protein species contained a lysine at codon 115 instead of arginine. The arginine 115 AGA codon is quite rare in *E. coli* genes. Misincorporation at other rare arginine codons (there are three others in the **a1** sequence) could not be ruled out, but the levels of lysine at other sites were too low for reliable detection by our ESI-MS methods.

The mass spectrum of **a1** homeodomain, expressed in *E. coli* grown in rich LB broth, revealed a lower level (approximately 5%) of (M-28) species compared with samples purified from minimal media preparations (Fig. 2B). To further investigate the dependence of misincorporation rate on media composition, we grew several protein preparations in minimal media containing varying levels of glucose. The mass spectra of all protein samples expressed in minimal medium, regardless of glucose content (varying from 2 g glucose/L to 10 g glucose/L), indicated that approximately 20–25% of the protein population contained misincorporated lysine (data not shown). Lower cell densities and slower growth curves are typically observed for *E. coli* cultures grown in minimal medium compared with cultures grown in rich medium, even though final **a1** homeodomain yields are comparable.

To block misincorporation at rare arginine codons, we set up two alternative expression systems. The first uses an alternate construct of the pCW/**a1**$_{66-126}$ plasmid (Baxter et al., 1994), pMF1, in which the R115 codon has been altered from 5′-AGA-3′ to the more common 5′-CGC-3′, using the Polymerase Chain Reaction–Splicing Overlap Extension (PCR-SOE) method (Horton, 1997). The second system co-expresses the *dnaY* gene plasmid, pUBS520 (Brinkmann et al., 1989), to produce the rare tRNA^{ARG} not normally abundant in *E. coli,* along with the **a1** homeodomain. Using both expression systems, the **a1** protein was expressed in the prototrophic K38 *E. coli* strain (Strain #4620, *E. coli* Genetic Stock Center) grown in minimal medium, purified and analyzed by mass spectroscopy for purity. The overall growth rates were slower for the co-expression system than for the K38/pMF1 cultures or the K38/pCW/al$_{66-126}$ control cultures. Although the growth rates of

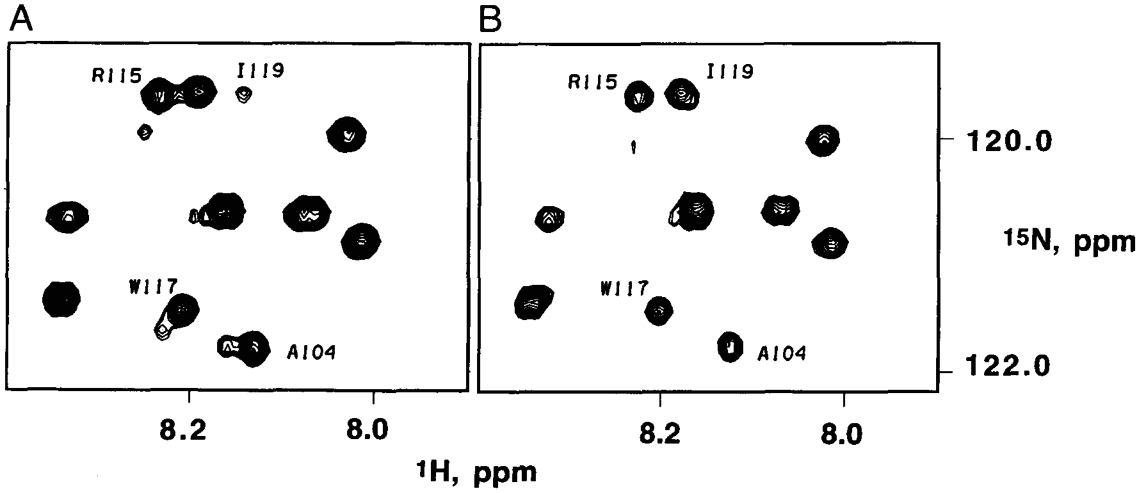

Fig. 1. Expansions of 500 MHz ^{1}H-^{15}N HSQC NMR spectra of recombinant **a1** homeodomain (**a1**$_{66-126}$) in 90% H_2O/10% D_2O, 25 mM deuterated acetate, pH 4.5, 100 mM KCl, 0.01% NaN_3 at 25 °C. **A:** 1.5 mM uniformly [^{15}N]-labeled **a1** homeodomain, expressed in minimal medium containing 10 g glucose/L, using the pCW/**a1**$_{66-126}$ overexpression system. **B:** 1.0 mM uniformly [^{15}N]-labeled **a1** homeodomain expressed in minimal medium containing 10 g glucose/L, using the pCW/**a1**$_{66-126}$/*dnaY* overexpression system. Several resonances are clearly doubled in **A** with ratios roughly 85%/15% major/minor components. Spectrum **B** does not include double peaks and resonance positions are at expected frequencies for correctly expressed protein.

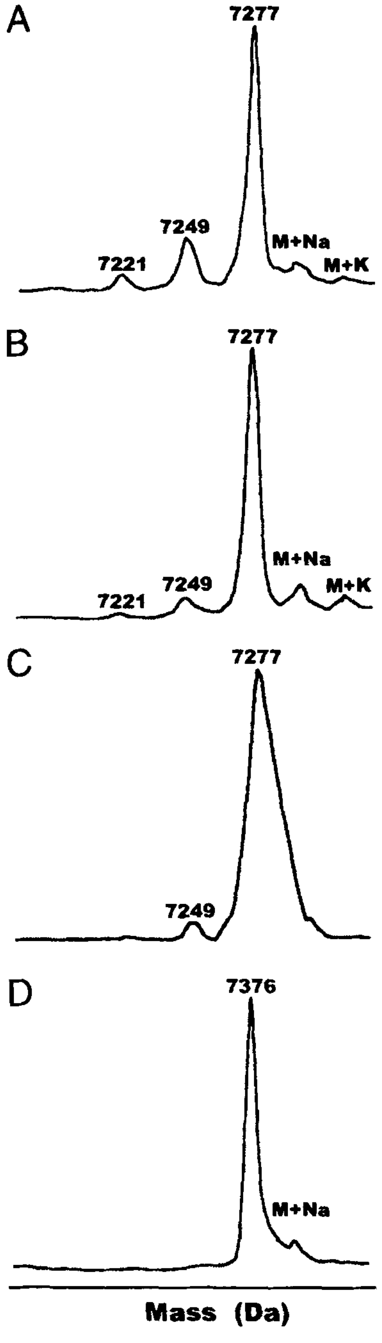

Fig. 2. Deconvoluted electrospray mass spectra of several preparations of the **a**1 homeodomain (**a**1$_{66-126}$). **A**: **a**1 homeodomain, coded with wild type sequence, expressed in minimal medium. **B**: **a**1 homeodomain, coded with wild type sequence, expressed in LB medium. **C**: **a**1 homeodomain, coded with the pMF1 plasmid containing the 115 codon replaced with GCG, expressed in minimal medium containing 10 g glucose/L. **D**: Uniformly ^{15}N-labeled **a**1 homeodomain, coded with wild type sequence co-expressed with *dnaY* plasmid, from *E. coli* grown in minimal medium containing 10 g glucose/L. Major and minor components are labeled with masses (Da). Higher mass components corresponding to sodiated and potassiated species are labeled. Co-expression of the **a**1 homeodomain with the *dnaY* plasmid results in homogenous protein preparations. Deconvoluted mass spectra of these protein preparations are free of minor components, (M-28) and (M-56), due to lysine misincorporation.

the two new expression systems were quite different, both expression systems resulted in similar final cell densities and similar amounts of purified protein. The protein yields were comparable to those obtained by previous methods (about 8 mg/L).

Mass spectroscopic analysis of the pMF1-encoded and purified **a**1$_{66-126}$ protein reveals that approximately 5% of the protein population has the −28 Da mass modification (Fig. 2C). Codon replacement at site 115 has reduced the level of misincorporation more than 50% in comparison with the previously expressed pCW/**a**1$_{66-126}$ protein. Mass analysis of the purified, uniformly ^{15}N labeled **a**1 protein, encoded by pCW**a**1$_{66-126}$/pUBS520 (*dnaY*), determined the sample to be >99% pure with the expected mass of 7,376 Da (Fig. 2D). Unlabeled **a**1 homeodomain, co-expressed with *dnaY*, was also analyzed and found to be >99% pure with expected mass of 7,277 Da (data not shown). Figure 1B shows an expansion of the ^{1}H-^{15}N HSQC spectrum of the **a**1 homeodomain co-expressed with the *dnaY* plasmid. The NMR resonances are no longer doubled. Co-expression of the *dnaY* plasmid completely blocks mistranslation at rare arginine codons in **a**1 homeodomain and allows production of a correctly expressed protein for structural and dynamic studies.

The mechanisms underlying the high rate of context-dependent lysine for arginine misincorporation observed in the **a**1 homeodomain preparations are still unclear. Our results, in combination with other studies, demonstrate that translation errors underlying misincorporation depend on the overexpression system. We find that protein expression in *E. coli* grown in minimal medium, results in higher rates of lysine for arginine misincorporation than *E. coli* grown in rich medium. In contrast, Day and coworkers (Day et al., 1996) reported no clear correlation between misincorporation and media composition for the human TSG-6 protein expression system. Other groups (Calderone et al., 1996; Day et al., 1996), recently reporting high rates of lysine misincorporation, used the T7 expression system (Studier et al., 1990) and produced protein in refractory inclusion bodies. In contrast, the **a**1 homeodomain is expressed in soluble form at lower levels of expression, albeit sufficient for structural studies. By co-expressing rare tRNA^{ARG} we have slowed the growth rate, compared with previous pCW-based **a**1 overexpression systems, but have now produced correctly expressed protein.

We observed a marked context dependence of lysine misincorporation during **a**1 homeodomain production. Over half the protein representing a single misincorporation during translation contained a lysine at codon 115. Rare AGA codons are found at sites 90, 91, 115, and 124 in the **a**1 coding sequence. Rosenberg and coworkers (Rosenberg et al., 1993) proposed that consecutive AGA codons cause translation errors. Although sequential AGA codons are found at sites 90 and 91 of the **a**1 code, these are not the major sites of misincorporation. Chen and Inouye (1990; 1994) have proposed that rare codons close to the initiator site (within the first 25 codons) modulate gene expression, leading to reduced levels of gene expression. Two groups (Seetharam et al., 1988; Day et al., 1996) have found lysine misincorporation at the first of several AGA-encoded arginines and suggest that rare AGA codons early in the sequence lead to preferential misincorporation at those sites. Arginine sites 90 and 91 in the **a**1 homeodomain coding sequence are coded by AGA sequences, but they lie just outside the first 25 codons of the sequence. Interestingly, arginine 115 is the major site of misincorporation, not R124, the last rare AGA codon in the sequence. The only distinguishing feature of the Arg 115 site is that it is flanked by two valine codons (GTA and GTT). Our experience with **a1** homeodomain expression suggests that misincorporation is not solely dependent on the concentration of tRNA^{ARG} but also depends on the context of the site.

Mass spectroscopic analysis of eukaryotic proteins, overexpressed for structural studies, should always be carried out to confirm correct translation. We want to highlight the importance of mass spectroscopic analysis of protein at various stages of expression system development. Researchers expressing proteins for NMR analysis, especially, know that it is not always straightforward to express proteins in minimal medium. The **a1** homeodomain expressed in rich culture medium contained low levels of misincorporation (<5%), but the same plasmid vector and *E. coli* strain grown in both minimal media and synthetic rich media (Baxter et al., 1994) results in >20% misincorporation. Even low level mistranslation rates (5–10%) are detrimental to high resolution structural studies since broadened or doubled NMR resonances complicate spectra and may wrongly suggest slow exchange dynamics. Only residues close in three-dimensional space to the site of the conservative lysine for arginine misincorporation appear in NMR spectra as "doubled peaks," much like the effect on spectra seen for protein samples containing differentially modified N-terminal methionines (Smith et al., 1997). We have found these problems and subtleties can be avoided in a straightforward, time-saving, and cost-effective manner by co-expressing rare tRNA^{ARG} with the eukaryotic protein of interest.

Electronic supplementary material: Plasmid construction, **a1** homeodomain expression, purification, and other experimental procedures are found in the Electronic Appendix.

Acknowledgments: We thank Dr. Arthur Palmer for helpful discussions on protein heterogeneity and Dr. Peter Buckel, Boehringer Mannheim, for generously supplying the pUBS520 plasmid (*dnaY*). We thank Stuart Lehrman for help with graphics. We gratefully acknowledge the use of the Wadsworth Center NMR Structural Biology and Mass Spectroscopy Core Facilities. This work was supported by NIGMS grant GM55361 to SMB.

References

Baxter SM, Gontrum DM, Phillips CL, Roth AF, Dahlquist FW. 1994. Heterodimerization of the yeast homeodomain transcriptional regulators $\alpha2$ and **a1**: Secondary structure determination of the a1 homeodomain and changes produced by $\alpha2$ interactions. *Biochemistry 33*:15309–15320.

Brinkmann U, Mattes RE, Buckel P. 1989. High-level expression of recombinant genes in Escherichia coli is dependent on the availability of the dnaY gene product. *Gene 85*:109–114.

Calderone TL, Stevens RD, Oas TG. 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.

Chen GF, Inouye M. 1990. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Res 18*:1465–1473.

Chen GT, Inouye M. 1994. Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. *Genes Dev 8*:2641–2652.

Day AJ, Aplin RT, Willis AC. 1996. Overexpression, purification, and refolding of link module from human TSG-6 in *Escherichia coli*: Effect of temperature, media, and mutagenesis on lysine misincorporation at arginine AGA codons. *Protein Expr Purif 8*:1–16.

Garcia GM, Mar PK, Mullin DA, Walker JR, Prather NE. 1986. The *E. coli* dnaY gene encodes an arginine transfer RNA. *Cell 45*:453–459.

Horton RM. 1997. In vitro recombination and mutagenesis of DNA. SOEing together tailor-made genes. *Methods Mol Biol 67*:141–149.

Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr Opin Biotechnol 6*:494–500.

Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G. 1993. Effects of consecutive AGG codons on translation in *Escherichia coli*, demonstrated with a versatile codon test system. *J Bacteriol 175*:716–722.

Seetharam R, Heeren RA, Wong EY, Braford SR, Klein BK, Aykent S, Kotts CE, Mathis KJ, Bishop BF, Jennings MJ. 1988. Mistranslation in IGF-1 during over-expression of the protein in *Escherichia coli* using a synthetic gene containing low frequency codons. *Biochem Biophys Res Commun 155*:518–523.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res 14*:5125–5143.

Smith SP, Barber KR, Shaw GS. 1997. Identification and structural influence of a differentially modified N-terminal methionine in human S100b. *Protein Sci 6*:1110–1113.

Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol 185*: 60–89. .

Wada K, Wada Y, Ishibashi F, Gojobori T, Ikemura T. 1992. Codon usage tabulated from the GenBank genetic sequence data. *Nucleic Acids Res 20*: 2111–2118.

Zahn K. 1996. Overexpression of an mRNA dependent on rare codons inhibits protein synthesis and cell growth. *J Bacteriol 178*:2926–2933.