BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

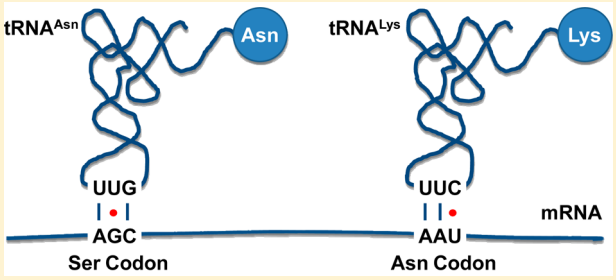
# G/U and Certain Wobble Position Mismatches as Possible Main Causes of Amino Acid Misincorporations

Zhongqi Zhang,* Bhavana Shah, and Pavel V. Bondarenko

Process and Product Development, Amgen Inc., Thousand Oaks, California 91320, United States

Supporting Information

**ABSTRACT:** A mass spectrometry-based method was developed to measure amino acid substitutions directly in proteins down to a level of 0.001%. When applied to recombinant proteins expressed in *Escherichia coli*, monoclonal antibodies expressed in mammalian cells, and human serum albumin purified from three human subjects, the method revealed a large number of amino acid misincorporations at levels of 0.001−0.1%. The detected misincorporations were not random but involved a single-base difference between the codons of the corresponding amino acids. The most frequent base differences included a change from G to A, corresponding to a G(mRNA)/U(tRNA) base pair mismatch during translation. We concluded that under balanced nutrients, G(mRNA)/U(tRNA) mismatches at any of the three codon positions and certain additional wobble position mismatches (C/U and/or U/U) are the main causes of amino acid misincorporations. The hypothesis was tested experimentally by monitoring the levels of misincorporation at several amino acid sites encoded by different codons, when a protein with the same amino acid sequence was expressed in *E. coli* using 13 different DNA sequences. The observed levels of misincorporation were different for different codons and agreed with the predicted levels. Other less frequent misincorporations may occur due to G(DNA)/U(mRNA) mismatch during transcription, mRNA editing, U(mRNA)/G(tRNA) mismatch during translation, and tRNA mischarging.




Protein sequence errors can cause protein misfolding, aggregation, and cell death.[1,2] The erroneous amino acid substitutions, instigated by either DNA mutations or errors during transcription or translation, are also a concern during the development of recombinant therapeutic proteins in the biopharmaceutical industry. Synthesized proteins typically follow the genetic information closely with only minor error frequencies, because of extensive editing and proofreading during protein synthesis.[3−7] Amino acid misincorporations, the main contributor to erroneous protein sequences, were estimated to occur during translation on a level of $10^{-5}$ to $10^{-3}$ per translated codon.[1,2,8−14] Even with such low levels of translational errors, when the fact that an average-length protein molecule is translated from hundreds of codons is taken into account, a high percentage of the protein molecules will contain at least one misincorporated amino acid. Although amino acid misincorporations occur naturally in proteins at very low levels, their frequency increases significantly upon translation from nonoptimal codons[15−17] or when cells are starved for certain amino acids.[1,18−20]

In most reported cases of amino acid substitutions, a limited number of them that occur at a high frequency were observed. Amino acid misincorporation, however, is a common phenomenon occurring in all cells, and therefore, a much larger number of low-abundance substitutions are expected. The lack of such information could be explained by the absence of large-scale and sensitive analytical techniques for the direct measurement of amino acid substitutions. Therefore, most of these substitutions were determined indirectly on a few selected special types. For example, a few amino acid misincorporations were monitored on the basis of the absence of a certain amino acid in a protein,[8] a change of charge,[9,10] or the activity[12,14] after the amino acid substitution.

Mass spectrometry (MS) and tandem mass spectrometry (MS/MS) have been widely used for protein structural characterization. Recent advances in high-resolution mass spectrometry make it possible to directly monitor very low levels of minor protein isoforms, including those caused by amino acid substitutions, on a large scale.[21] To monitor small amounts of protein isoforms on the protein level, a bottom-up liquid chromatography−tandem mass spectrometry (LC−MS/MS) approach, i.e., proteolytic digestion followed by LC−MS/MS analysis, is commonly accepted as the most sensitive technique. To detect an amino acid substitution, the substituted peptide must be identified on the basis of its determined mass and sequence information in the fragmentation pattern (MS/MS). This article describes our recently developed platform for large-scale screening for amino acid substitutions and results obtained when the platform was used to measure a large number of amino acid substitutions commonly occurring in recombinant as well as natural proteins. The methodology is based on high-

**Received:** July 25, 2013
**Revised:** September 10, 2013
**Published:** October 15, 2013

ACS Publications © 2013 American Chemical Society



dx.doi.org/10.1021/bi401002c | *Biochemistry* 2013, 52, 8165−8176

resolution mass spectrometry, combined with an advanced data acquisition process and data analysis algorithms developed in house. It is capable of detecting amino acid substitutions in recombinant proteins at levels approaching 0.001% ($10^{-5}$), comparable with the expected natural levels of translational and transcriptional errors. The mass spectrometry-based method greatly expands the range of amino acid substitutions that can be detected and allows direct measurement of the protein sequence errors on each amino acid residue.

## MATERIALS AND METHODS

**Materials.** Two IgG2 monoclonal antibodies (mAbs) were expressed and purified from Chinese hamster ovary (CHO) cells under variable cell culture conditions. Six recombinant proteins, including one small protein and five Fc fusion proteins, were expressed and purified from *Escherichia coli*. To test the proposed hypothesis, one of the Fc fusion proteins was expressed in *E. coli* using 13 constructs of different DNA sequences, among which codons for some of the amino acids were varied.[17] A total of 23 fermentation runs were performed using these constructs, and the purified protein from each run was analyzed. All recombinant proteins were produced at Amgen Inc.

To monitor amino acid misincorporations occurring in natural proteins, human serum albumin (HSA) was purified from three nonrelated human subjects using blue trisacryl M-affinity chromatography (ARVYS Proteins, Stamford, CT) to a purity of >99% as determined by densitometry on a Coomassie-stained sodium dodecyl sulfate−polyacrylamide gel electrophoresis gel.

**Proteolytic Digestion.** Proteolytic digestions of proteins were performed similarly as described by Ren et al.[22] with minor modifications. Briefly, each protein sample was denatured and disulfide-reduced in a Tris buffer (pH 7.5) containing 7.2 M guanidine hydrochloride and 6 mM dithiothreitol (30 min at 37 °C), followed by alkylation with 14 mM iodoacetic acid or iodoacetamide. The reduced and alkylated protein (at 0.2−1.0 mg/mL) was buffer-exchanged into 0.1 M Tris buffer (pH 7.5) using ultrafiltration on a 10 kDa cutoff membrane (Vivaspin 500, Sartorius Stedim Biotec) or a buffer-exchange spin column (Pierce detergent removing spin column) according to the manufacturer's suggested procedure, followed by addition of an appropriate amount of trypsin (Roche) or Lys-C (Wako) to achieve a substrate:enzyme ratio of 20:1 and incubation at 37 °C for 1 h (trypsin) or 2 h (Lys-C). The digestion was quenched by adding acetic acid to decrease the pH to approximately 5.

**LC−MS/MS Analysis.** Tryptic digests of IgG2 mAbs expressed in CHO cells were analyzed on a Thermo Scientific (San Jose, CA) LTQ-Orbitrap high-resolution mass spectrometer connected to an Agilent (Santa Clara, CA) 1290 Infinity LC system. A Waters (Milford, MA) BEH 300 C18 reversed-phase column (1.7 $\mu$m particle, 150 mm × 2.1 mm) at 65 °C was used for the separation, followed by electrospray ionization. Peptides were eluted with an acetonitrile gradient (1 to 40% over 90 min, followed by column washing with 40 to 99% acetonitrile) at a flow rate of 0.2 mL/min, with 0.02% trifluoroacetic acid (TFA) in the mobile phase. Approximately 18 $\mu$g of protein digest was injected for each analysis.

Tryptic digests of proteins expressed in *E. coli* were analyzed on the Thermo Scientific LTQ-Orbitrap mass spectrometer connected to an Agilent 1200SL LC system. A Waters BEH300 C-18 column (1.7 $\mu$m particle, 100 mm × 2.1 mm) at 50 °C was used for the separation, followed by electrospray ionization. Peptides were eluted with an acetonitrile gradient (0.5 to 20% over 20 min and then 20 to 35% over 40 min, followed by column washing from 35 to 99% acetonitrile) at a flow rate of 0.3 mL/min, with 0.04% TFA in the mobile phase. Approximately 25 $\mu$g of protein digest was injected for each analysis.

Lys-C digests of HSA purified from human sera were analyzed on a Thermo Scientific LTQ-Orbitrap Elite high-resolution mass spectrometer connected to an Agilent 1290 Infinity LC system. A Waters BEH 300 C18 reversed-phase column (1.7 $\mu$m particle, 150 mm × 2.1 mm) at 65 °C was used for the separation, followed by electrospray ionization. Peptides were eluted with an acetonitrile gradient (0.5 to 40% over 120 min, followed by column washing from 40 to 99% acetonitrile) at a flow rate of 0.2 mL/min, with 0.1% formic acid in the mobile phase. Approximately 18 $\mu$g of protein digest was injected for each analysis.

The LTQ-Orbitrap instrument was set up to collect one full-scan spectrum at a resolution of 60000 (at $m/z$ 400), followed by two or three data-dependent MS/MS spectra of the most abundant ions in the linear trap. The LTQ-Orbitrap Elite instrument was set up to collect one full-scan spectrum at a resolution of 120000, followed by five data-dependent MS/MS spectra of the most abundant ions in the linear trap. MS/MS spectra were collected with dynamic exclusion, using collision-induced dissociation with a normalized collision energy of 35%. For all mAb and HSA samples, automated precursor ion exclusion (PIE),[23] as described briefly below, was used to acquire a large number of unique, high-quality MS/MS spectra. For HSA analysis on the LTQ-Orbitrap Elite instrument, the automatic gain control target value for the full scan was set to a high value of $2 \times 10^6$ to improve the detection of low-abundance ions.

**Precursor Ion Exclusion.** Dynamic exclusion is often used to acquire as many unique MS/MS spectra as possible for the purpose of identifying a large number of peptides. However, dynamic exclusion is limited to within a single LC−MS/MS run. For samples with similar compositions, however, after the MS/MS spectra of an ion are acquired in one run, the MS/MS spectra of the same ion in the later runs do not need to be acquired, because the identities of these ions can be derived from the MS/MS spectra of the same ion in the previous run. This scheme can be realized by analyzing a data file to generate a "precursor ion exclusion list" from the ions that has triggered MS/MS and then adding the "exclusion list" to the MS method file, which will be used in later runs. This PIE approach is implemented in MassAnalyzer for fully automated execution. In this approach, after each the LC−MS/MS run is finished, the raw data file is automatically analyzed to generate a list of ions to be rejected, and this list is applied to the following run as a rejection list for selection of MS/MS precursor ions. The list contains information regarding the $m/z$ and elution time range of each ion. The PIE list is generated in a way so that the next runs will generate as many unique and high-quality MS/MS spectra as possible. Automated PIE greatly increases the number of unique MS/MS spectra acquired and therefore significantly increases the number of identified peptides.[23]

**Data Analysis.** Analysis of LC-MS/MS data was performed on MassAnalyzer.[21] MassAnalyzer identifies peptides (including modified peptides and peptides with amino acid substitutions) by comparing the experimental MS/MS data to theoretically predicted MS/MS spectra based on custom-developed empirical kinetic models of peptide fragmentation.[24−27] Comparing the experimental spectrum to the accurately predicted theoretical spectrum improves the reliability of peptide identification as compared to the reliability of conventional methods that do not consider fragment ion intensities.[21] When identifying peptides

with amino acid substitutions, MassAnalyzer first identifies a candidate by comparing the determined mass of an unknown peptide to those of all identified peptides. If the mass difference corresponds to a potential substitution of one of the residues in the peptide with another residue, then the predicted MS/MS spectrum of the substituted peptide will be generated and compared to the experimental MS/MS spectrum. A match between the two fragmentation spectra[21] indicates a positive identification of an amino acid substitution. An example of peptide identification by MassAnalyzer is shown in Figure S1 of the Supporting Information.

Levels of amino acid substitutions are automatically quantified by MassAnalyzer, together with all other identified modifications, using MS ion intensities (peak areas under the selected ion chromatograms constructed using a matched window function).[21] MassAnalyzer outputs a table of amino acid substitutions and other chemical modifications, with quantity levels estimated from their MS intensities. After automated retention time alignment,[28] all samples analyzed together are automatically compiled in the same table. A peptide needs to be identified in only one sample, and then its full MS scan ion features (chromatographic peak profiles of all isotopes) are utilized for quantifications in all samples.

**Distinguishing Amino Acid Substitutions from Post-Translational Modifications.** Because amino acid substitutions are rare compared to post-translational and other chemical modifications, some chemical modifications with the same mass changes as amino acid substitutions are frequently misinterpreted as substitutions, even with the 2−8 ppm mass accuracy of the mass spectrometer used in this work. The major challenge of the described methodology is to distinguish amino acid substitutions from a large number of possible chemical modifications in the protein.

The misinterpretations can be caused by either a modification of the same residue or a modification of a nearby residue, but without sufficient sequence coverage in MS/MS to determine the exact location of the modification. Ambiguity sometimes remains even when the site of modification is identified precisely. For example, hydroxyl radical-induced oxidation may happen on virtually all 20 amino acids.[29,30] Oxidation of Ala and Phe will be interpreted as A → S (representing that A is replaced by an S) and F → Y substitutions, respectively. Oxidation of Ala and Phe, unfortunately, cannot be confidently distinguished from A → S and F → Y substitutions. Other examples include N → D (deamidation of Asn), Q → E (deamidation of Gln), H → N (oxidation of histidine[31]), etc. To remove the false positives, we have tabulated them and established a set of simple rules to follow to distinguish true amino acid substitutions from chemical modifications.

First, we compare the MS/MS spectrum of each modified (including the amino acid-substituted form) peptide to that of the unmodified peptide to determine whether the mass change is indeed on the substituted residue, and whether the mass change can be explained by a known modification, including modifications occurring during sample storage, sample preparation, and LC−MS analysis. Table S1 of the Supporting Information shows common modifications with mass changes similar to those of some substitutions that may cause false positives or false negatives. To maintain the high level of confidence in the identified amino acid substitutions, if a mass change could be interpreted as a chemical modification, then it was not considered as an amino acid substitution. Attention was also given to gas-phase reactions, which produce a modified peptide with the same retention time as the unmodified peptide. If any modifications are known to exist in the samples (e.g., carbamylation, formylation, etc.), they are added to the list of modifications for which to search to reduce the number of false positives. Although manual removal of false positives required extensive knowledge and experience in interpreting peptide MS/MS data initially, with the help of Table S1 of the Supporting Information, it becomes much less labor-intensive. At the end, many of these known modifications have also been incorporated into MassAnalyzer for automated evaluation to reduce the number of false positives.

If MS/MS does not provide enough information about the location of the mass change, we then check whether the same mass change is observed on other peptides, in which the modified residue can be determined from their MS/MS data. Modifications and amino acid substitutions caused by misincorporations (but not DNA mutations) usually happen on the same residue at different locations. To be considered as an amino acid misincorporation, the same type of substitution must be observed on multiple sites of the same molecule or other molecules produced by a similar process.

It frequently happens that the available information in the MS/MS spectra is not sufficient to distinguish an amino acid substitution from a chemical modification. Any supporting evidence will be useful for determining whether the substitution proposed by MassAnalyzer is a true substitution. A frequently used rule is related to the specificity of the protease used for the digestion. Some amino acid substitutions involve a residue that matches the specificity of the protease. These substitutions can often be confirmed by whether they are cleaved by the protease. Examples include G to D for Asp-N, G to E for Glu-C, N to K for trypsin and Lys-C, S to R and R to Q for trypsin, etc. However, these residues can also be added to the N-terminus or C-terminus of a proteolytic peptide through protease-catalyzed transpeptidation during digestion;[32,33] care must be taken to eliminate these artifacts. When multiple analyses were performed using complementary proteases, an alternative peptide containing the same residue may provide further evidence or cross confirmation of the identified amino acid substitutions. The elution order of a peptide with substitution relative to the wild-type peptide is yet another piece of supporting information confirming the identification, based on the change in the hydrophobicity of the substitution.

By including only misincorporations that are confidently identified, we might exclude some true misincorporations. However, it is reasonable to assume that the excluded misincorporations are random in nature with regard to the misincorporation mechanisms and, therefore, would not introduce significant bias into the result. Although some misincorporations cannot not be detected because of their coincidence with common chemical modifications, considering there are 19 × 19 = 361 types of possible misincorporations in total, they should not significantly alter the statistics.

**Distinguishing DNA Mutations from Amino Acid Misincorporations.** DNA mutations are rare, random, and mostly independent events. As a result, the chance of the same mutation happening in multiple independent cell lines is extremely small. Amino acid misincorporations, on the other hand, depend on the cell line to a much lesser degree. DNA mutations are also location-dependent; i.e., the same mutation would be very unlikely to appear on the same amino acid at a different location. Amino acid misincorporations, on the other hand, appear at multiple locations for amino acids encoded by

**Table 1. Levels of Amino Acid Substitutions Detected in IgG2 mAb (X) Produced from Six Transfected CHO Cell Clones under Different Cell Culture Conditions**

| | | | average level of substitution ($\times 10^{-5}$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 2[a] | 3[a] | 3[a] | 4[a] | 4[a] | 4[a] | 4[a] | 4[a] |
| | | | clone 2-18 | clone 3-9 | clone 3-9 | clone 4-11 | clone 4-1 | clone 4-7 | clone 4-11 | clone 4-14 |
| | codon | no. of sites[b] | B1[c] | A[c] | B2[c] | A[c] | B1[c] | B2[c] | B2[c] | B2[c] |
| misincorporations[d] | | | | | | | | | | |
| G → D | GGC | 5/17 | 30 | 17 | 10 | 16 | 12 | 12 | 9 | 9 |
| M → I | AUG | 3/10 | 5 | 7 | 5 | 5 | 4 | 3 | 3 | 4 |
| N → K | AAC | 4/22 | 5 | 3 | 11 | 20 | 14 | 21 | 20 | 24 |
| | AAU | 3/5 | 6 | 2 | 6 | 26 | 22 | 29 | 20 | 21 |
| N → S | AAC | 8/22 | 15 | 1 | 7 | 17 | 82 | 69 | 50 | 59 |
| | AAU | 4/5 | 13 | 2 | 7 | 17 | 84 | 64 | 50 | 55 |
| S → N | AGC | 5/29 | 29 | 41 | 22 | 29 | 21 | 21 | 19 | 17 |
| | AGU | 1/7 | 13 | 14 | 10 | 10 | 12 | 12 | 9 | 11 |
| S → R | AGC | 1/29 | 0 | 3 | 1 | 3 | 0 | 0 | 1 | 0 |
| | AGU | 1/7 | 3 | 3 | 1 | 4 | 2 | 2 | 2 | 2 |
| V → I | GUC | 4/25 | 7 | 19 | 7 | 9 | 5 | 5 | 6 | 5 |
| mutations[e] | | | | | | | | | | |
| HC:F136V | UUC | 1/21 | 0 | **47** | **55** | 0 | 0 | 0 | 0 | 0 |
| LC:T107S | ACC | 1/27 | 59 | 19 | 35 | **8620** | **8240** | **7320** | **7770** | **7420** |
| LC:N142K | AAU | 1/5 | **7450** | 2 | 3 | 8 | 11 | 13 | 12 | 10 |

[a]Pool of transfected cells. [b]Number of observed amino acid substitutions compared to the total number of specified codons for the protein. [c]Cell culture conditions differ in the concentrations of several nutrients, including the concentrations of amino acids. Conditions B1 and B2 are slight variations of the same condition B. [d]G → D indicates that the glycine residue is replaced by an aspartic acid residue. [e]Potential DNA mutations are shown in bold.

the same codon. In practice, to be considered as an amino acid misincorporation, the same type of substitution should be observed more than once at different locations, either on the same protein or on different proteins produced with a similar type of cell line and process. Therefore, if an amino acid substitution is observed in only one cell clone (or one human subject) and one location, then the substitution is most likely caused by DNA mutation. It is therefore important that multiple diverse clones are analyzed at the same time for confidently distinguishing DNA mutations and amino acid misincorporations.

## RESULTS AND DISCUSSION

**Amino Acid Misincorporations in Recombinant Proteins Expressed in Mammalian Cells.** Eight batches of IgG2 mAb (X) were produced from six transfected CHO cell clones and two different cell culture conditions. The two cell culture conditions differed slightly in the concentrations of nutrients, including the concentrations of amino acids. The purified mAb samples were digested with trypsin and analyzed by LC−MS/MS. The LC−MS/MS data were analyzed by MassAnalyzer with automated algorithms for identification and quantitation of amino acid substitutions. The fragmentation spectra of these peptides were manually examined to eliminate any false positives. After false positives had been removed, a large list (42 in total) of amino acid substitutions remained, with levels generally below $10^{-3}$ (Table 1).

Most of these substitutions were caused by amino acid misincorporations, because they appeared in many different locations of the protein and had similar abundances. Three exceptions, however, were found in the list. One was the light chain (LC) T107S substitution (~8%) found in clones 4-1, 4-7, 4-11, and 4-14. The second was the LC N142K substitution (~7%) found only in clone 2-18. The third was a very small amount (~0.05%) of heavy chain (HC) F136V substitution found only in clone 3-9. These three amino acid substitutions were most likely caused by DNA mutations. Although the codons used by LC T107 (ACC) and HC F136 (TTC) are common in the DNA sequence of the protein (occur 27 and 21 times, respectively), the T → S and F → V substitutions were observed at only one site in each molecule. N → K substitutions, however, were observed in many other sites of the same clone 2-18 and other clones with much lower levels, because it happened to be a common misincorporation. The high percentage of LC N142K potential mutation (~7%) in clone 2-18 clearly distinguishes it from misincorporations. The two S → R substitutions, although detected at only one site for each codon, are unlikely to be mutations because they are present in multiple independent clones with similar levels. It is interesting that the same T107S substitution appeared in all four clones derived from the same pool of transfected cells (pool 4), suggesting that the four parent cells, from which the four clones were derived, were from the same mutated cell. LC−MS/MS screening of multiple diverse clones provides an opportunity for selecting a clone without any or with an only minimal level of mutations. The final clone should be subjected to confirmational sequencing on a DNA or RNA level as a part of the clone selection routine. DNA and RNA sequencing was not performed here because DNA mutations were not the main focus of this study.

In addition to amino acid misincorporations observed in the IgG2 mAb shown in Table 1, misincorporations detected in another IgG2 mAb (Y) (expressed in a same CHO host cell line with two different cell culture conditions, A and B) are listed in Table 2. Although several types of misincorporation were on similar levels between the two process conditions, N → K, S → N, and S → R misincorporations were significantly different, indicating that they were process-dependent. The same types of misincorporation, at levels below $10^{-3}$, were generally detected in


dx.doi.org/10.1021/bi401002c | Biochemistry 2013, 52, 8165−8176

**Table 2. Average Levels ($10^{-5}$) of Amino Acid Misincorporations Detected in IgG2 mAb (Y) Expressed in the Same Transfected CHO Cell Line under Two Different Cell Culture Conditions (each with four biological replicates)**

| | | | average level of misincorporation ($\times 10^{-5}$) ± the standard deviation ($n = 4$) | |
|---|---|---|---|---|
| misincorporation | codon | no. of sites[b] | A[a] | B[a] |
| D → G | GAC | 2/19 | 1 ± 0 | 1 ± 0 |
| D → N | GAC | 4/19 | 3 ± 1 | 2 ± 0 |
| G → D | GGC | 8/25 | 19 ± 2 | 15 ± 2 |
| | GGU | 1/2 | 75 ± 45 | 77 ± 68 |
| M → I | AUG | 2/9 | 6 ± 0 | 7 ± 1 |
| N → K | AAC | 7/20 | 0 ± 0 | 5 ± 1 |
| | AAU | 3/6 | 1 ± 0 | 10 ± 2 |
| S → N | AGC | 10/29 | 60 ± 10 | 18 ± 1 |
| | AGU | 1/3 | 27 ± 5 | 9 ± 1 |
| S → R | AGC | 3/29 | 9 ± 1 | 4 ± 0 |
| | AGU | 2/3 | 5 ± 1 | 2 ± 0 |
| V → I | GUC | 4/19 | 23 ± 1 | 23 ± 3 |

[a]Cell culture condition A is a control condition. Condition B is a larger-scale process. [b]Number of observed amino acid misincorporations compared to the total number of specified codons for the protein.

IgG2 mAb (Y) (Table 2) as in IgG2 mAb (X) (Table 1). In addition, a few new low-abundance misincorporations were detected in IgG2 mAb (Y) at much lower levels [$<10^{-4}$ for D → G and D → N (Table 2)]. In general, the levels of amino acid misincorporations in the mammalian cells were on the order of $10^{-5}$ to $10^{-3}$, in agreement with the previously reported error rate for protein translation.

**Amino Acid Misincorporations in Recombinant Proteins Expressed in *E. coli*.** Amino acid misincorporations were assessed in several recombinant proteins expressed in *E. coli* (Table 3). All substitutions were determined to be caused by misincorporations instead of DNA mutations because the same substitutions were observed at several different sites or multiple proteins. The levels of these misincorporations were mostly on the order of $10^{-4}$ to $10^{-3}$ (0.01−0.1%).

**G/U Mismatches during Codon Recognition.** Amino acid misincorporation can occur through one of the three mechanisms: (1) transcriptional error, (2) mischarging of an (incorrect) amino acid to the tRNA, and (3) mRNA/tRNA mismatching. Evidence presented in this report, as discussed below, suggests that under tested cell culture conditions with sufficient amino acids as nutrients, most misincorporations occur through mRNA/tRNA mismatching, with a few exceptions that may be caused by transcriptional errors or tRNA mischarging caused by temporary amino acid starvation.

All confidently identified misincorporations (Tables 1−3) involved two amino acids with a single base difference between their codons. Table 4 summarizes the observed misincorporations in several CHO-expressed mAb molecules, including data shown in Tables 1 and 2 and additional data for other molecules. Substituted amino acid pairs in these molecules include D/N, G/D, M/I, N/K, N/S, S/R, and V/I.

As a summary of Table 3 for proteins expressed in *E. coli* and Table 4 for proteins expressed in CHO cells, a total of 20 substituted amino acid pairs were observed. For the purpose of generalizing some common rules governing the observed misincorporations, the base differences between the codons of the 20 substituted amino acid pairs are summarized in Table 5.

A striking feature revealed in Table 5 is that 45% of the observed misincorporations involve amino acids with a G → A base change in their codons. In fact, of 15 possible misincorporations involving G → A base changes, nine of them (60%) were observed in this work. It needs to be taken into account that of 150 possible misincorporations involving any single-base changes, only 20 of them (13%) were observed here. Take another calculation for example; in the two mAbs shown in Tables 1 and 2, a total of 93 Ser residues are encoded by UCX and 68 are encoded by AGX, but all 17 observed S → N substitutions are from Ser encoded by AGX (corresponding to a G → A base change), with chance of coincidence calculated to be $1.2 \times 10^{-7}$. The frequent observation of G → A base changes can be explained by a $G^{mRNA}/U^{tRNA}$ base pair mismatch during translation, with G in the mRNA codon and U in the tRNA anticodon (Figure 1B). The statistical bias in the misincorporations revealed in this study indicates that the G/U mismatch is the most frequent mismatch during codon recognition. G/U mismatches are well-known to occur frequently in nucleic acid secondary structures[34−37] and mRNA/tRNA interactions[13,38] due to their similar binding energies as conventional Watson−Crick base pairs.[13,39] The G/U mismatch is also frequently observed at the wobble codon position (the third base) during translation.[40] A G/U mismatch in the first and second base during codon recognition was also proposed as early as 1963.[41] The near-cognate $G^{mRNA}/U^{tRNA}$ mismatch in the first base position during codon recognition was also experimentally found to have a binding affinity very similar to that of the cognate $A^{mRNA}/U^{tRNA}$ match.[38] These prior reports strongly support the notion that the uncovered misincorporations involving G → A base changes are caused by a competition between the near-cognate $G^{mRNA}/U^{tRNA}$ mismatch and the cognate $G^{mRNA}/C^{tRNA}$ match during codon recognition (Figure 1B). The favorable G → A base changes in the codons during amino acid misincorporation are also supported by several other previous observations, including G → E, G → D, R → K, and R → Q misincorporations observed in other recombinant proteins expressed in *E. coli*.[14,15,17,42,43] Our conclusion is also supported by the work of Kramer and Farabaugh that showed translational errors were largely determined by tRNA competition. When the cognate $tRNA^{Arg}$ was overexpressed in *E. coli*, the level of R → K misincorporation was significantly reduced.[14] This observation supports our claim that R → K misincorporation is caused by the competition between arginyl-$tRNA^{Arg}$ (cognate match) and lysyl-$tRNA^{Lys}$ (G/U mismatch at the second codon position).

**Wobble Position Mismatches during Codon Recognition.** The wobble codon position is believed to be less stringent in base pair matching. Indeed, a base change in the third position explains an additional 25% (Table 5) of misincorporations, making a total of 70% of misincorporations explained by either the favorable $G^{mRNA}/U^{tRNA}$ mismatches or wobble position mismatches. Of 14 possible misincorporations involving mismatches in their third codon positions, six of them (including a G → A change at the third position) are observed in this work.

Among the six types of misincorporations caused by mismatches at the third codon position during codon recognition, one of them is a G/U mismatch, four are C/U or U/U mismatches, and the last one is an A/G mismatch (Table 5). Although they appear less frequently, A/G, C/U, and U/U mismatches are also observed in RNA secondary structures.[36] Because of the less stringent nature of the third codon position


dx.doi.org/10.1021/bi401002c | *Biochemistry* 2013, 52, 8165−8176

**Table 3. Detected Levels of Misincorporation in Several Recombinant Proteins Expressed in *E. coli*[a]**

| | | | possible cause | | average level of misincorporation ($\times 10^{-5}$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| misincorporation | codon | base difference | transcriptional mismatch | translational mismatch[b] | 1 (20 kDa) | 2 (46 kDa) | 3 (46 kDa) | 4 (33 kDa) | 5 (29 kDa) | 6 (42 kDa) |
| A → T | GCC | G1 → A | | G1/U | 29(1) | | | | | |
| | GCU | | | | | | | | | 28(1) |
| D → E | GAC | C3 → A or G | | C3/U | 44(3) | 27(8) | 39(9) | 30(2) | 39(2) | 18(5) |
| | GAU | U3 → A or G | | U3/U or U3/C | 70(3) | 42(2) | 39(2) | | 186(2) | |
| E → K | GAA | G1 → A | | G1/U | | 2(1) | 30(1) | 5(1) | 6(1) | |
| G → D | GGC | G2 → A | | G2/U | 52(2) | 51(3) | 80(3) | 17(1) | 5(1) | |
| | GGU | | | | | 18(1) | 39(1) | | | |
| G → E | GGA | G2 → A | | G2/U | 187(3) | 304(8) | 82(3) | | 148(1) | 18(1) |
| | GGG | | | | 6(1) | 12(3) | 29(3) | 80(1) | 6(1) | |
| H → Q | CAC | C3 → A or G | | C3/U | 24(2) | | 12(1) | | | |
| | CAU | U3 → A or G | | U3/U or U3/C | 83(1) | 74(1) | 60(1) | | | |
| M → I[c] | AUG | G3 → A, etc. | | G3/U | 4(1) | 1(1) | | 4(1) | 9(1) | 15(1) |
| M → T | AUG | U2 → C | | U2/G | 5(1) | 30(2) | 54(2) | | | 20(2) |
| N → K | AAC | C3 → A or G | | C3/U | | 3(1) | 3(1) | | | 10(2) |
| | AAU | U3 → A or G | | U3/U or U3/C | 48(1) | 9(4) | 10(4) | 14(2) | 38(4) | 27(3) |
| P → L | CCC | C2 → U | G/U | | 23(2) | 21(3) | 25(3) | 10(1) | 10(1) | 6(2) |
| P → S | CCA | C1 → U | G/U | | 6(1) | | | | | |
| | CCC | | | | 9(1) | | | | | |
| | CCG | | | | 11(1) | | | | | |
| Q → H | CAA | A3 → C or U | | A3/G | 9(3) | | | | | |
| R → Q | CGA | G2 → A | | G2/U | 34(1) | 31(1) | 24(1) | | | |
| | CGG | | | | 239(5) | 127(5) | 30(2) | 159(3) | 406(3) | 386(3) |
| S → N | AGC | G2 → A | | G2/U | | 52(7) | 43(7) | 21(6) | 107(7) | 44(8) |
| | AGU | | | | | | | | | 6(1) |
| V → I[d] | GUC | G1 → A | | G1/U | 40(1) | 29(8) | 40(8) | 24(6) | 44(6) | 29(7) |
| | GUU | | | | | 2(1) | 4(2) | | | 15(1) |
| Y → N | UAC | U1 → A | | U1/U | | 30(4) | 29(4) | 8(3) | 28(3) | 27(4) |

[a]The levels shown are the average determined levels of all detected misincorporations (number of detected misincorporations shown in parentheses) on amino acids with the same codon. Numbers 1−6 represent six different recombinant proteins. Proteins 2−6 are Fc fusion proteins, with the mass value representing one of the two identical chains. [b]A base pair mismatch is represented as X*n*/Y, where X is the base on the mRNA codon, Y is the base on the tRNA anticodon, and *n* is the position of the mismatched base in the codon. [c]I and L are indistinguishable, but the M → I misincorporation is more likely than the M → L misincorporation because the former involves a favorable wobble position G/U mismatch. [d]The V → I misincorporation is more likely than the V → L misincorporation because it involves a favorable G/U mismatch.

**Table 4. Amino Acid Misincorporations Detected in Several CHO-Expressed Proteins**

| | | | possible cause | | |
|---|---|---|---|---|---|
| misincorporation | codon difference | base difference | transcription | translation | error rate[a] |
| D → G | GAC → GGC | A2 → G | G/T mismatch or mRNA editing | | $6 \times 10^{-6}$ |
| D → N | GAC → AAC | G1 → A | | G1/U mismatch | $2 \times 10^{-5}$ |
| G → D | GGC → GAC | G2 → A | | G2/U mismatch | $2 \times 10^{-4}$ |
| | GGU → GAU | G2 → A | | | $8 \times 10^{-4}$ |
| M → I[b] | AUG → AUA | G3 → A | | G3/U mismatch | $5 \times 10^{-5}$ |
| N → K | AAC → AAA or AAG | C3 → A or G | | C3/U mismatch | $9 \times 10^{-5}$ |
| | AAU → AAG or AAA | U3 → A or G | | U3/U or U3/C mismatch | $1 \times 10^{-4}$ |
| N → S | AAC → AGC | A2 → G | G/T mismatch or mRNA editing | tRNA mischarge | $4 \times 10^{-4}$ |
| | AAU → AGU | | | | $4 \times 10^{-4}$ |
| S → N | AGC → AAC | G2 → A | | G2/U mismatch | $3 \times 10^{-4}$ |
| | AGU → AAU | | | | $1 \times 10^{-4}$ |
| S → R | AGC → AGA or AGG | C3 → A or G | | C3/U mismatch | $4 \times 10^{-5}$ |
| | AGU → AGA or AGG | U3 → A or G | | U3/U or U3/C mismatch | $3 \times 10^{-5}$ |
| V → I[c] | GUC → AUC | G1 → A | | G1/U mismatch | $2 \times 10^{-4}$ |

[a]The error rate is the average abundance of the misincorporations detected in all molecules in this study. [b]I and L are indistinguishable, but the M → I misincorporation is more likely than the M → L misincorporation because it involves a favorable wobble position G/U mismatch. [c]The V → I misincorporation is more likely than the V → L misincorporation because it involves a favorable G/U mismatch.

during codon recognition, these less frequent mismatches become significant during codon recognition. Of seven possible misincorporations involving C/U or U/U mismatches in the third codon position, four of them are observed here. However, of six possible misincorporations involving an A/G mismatch in the third codon position, only one is observed, indicating A/G


dx.doi.org/10.1021/bi401002c | *Biochemistry* 2013, 52, 8165−8176

**Table 5. Summary of Base Differences in the Codons of Substituted Amino Acids in Recombinant Proteins Expressed in *E. coli* and CHO Cells[a]**

| base difference | no. of amino acid pairs[b] | fraction (%) (of a total of 20) | possible cause | |
|---|---|---|---|---|
| | | | transcription | translation |
| G → A | 9/15 | 45 | | G/U mismatch |
| third base C or U → A or G | 4/7 | 20 | | C/U or U/U mismatch |
| third base A → C | 1/6 | 5 | | A/G mismatch |
| C → U | 2/11 | 10 | G/U mismatch (*E. coli* only) | |
| A → G | 2/14 | 10 | T/G mismatch or mRNA editing | |
| U → C | 1/11 | 5 | | U/G mismatch |
| U → A | 1/14 | 5 | | U/U mismatch |
| total | 20/77 | 100 | | |

[a]The most probable base changes are given in the first two rows. [b]Number of observed amino acid pairs compared with the total number of possible pairs with the specified base change.

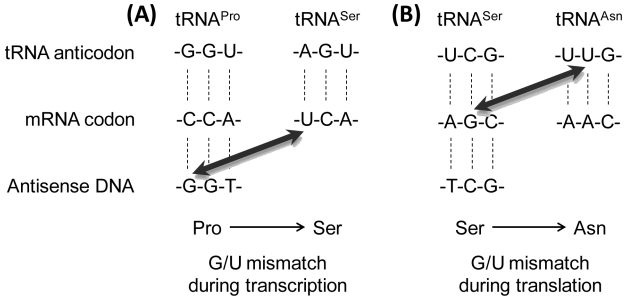



**Figure 1.** (A) Amino acid misincorporations involving a C → U base change in their mRNA codons can be explained by a $G^{DNA}/U^{mRNA}$ base pair mismatch (indicted by the double arrow) during transcription. (B) Amino acid misincorporations involving a G → A base change can be explained by a $G^{mRNA}/U^{tRNA}$ base pair mismatch during translation.

mismatches in the third codon position are not as frequent as C/U or U/U mismatches. Note that the wobble positions in tRNA molecules are sometimes modified for pairing with degenerate codons; this further increases the chance of mismatches in the wobble position.

**Other Mismatches during Codon Recognition.** The fact that U/C and U/U mismatches are also observed in RNA secondary structures[36] supports the observation of Y → N misincorporation (U/U mismatch) in *E. coli* (Table 3). Most of the early discovered amino acid misincorporations in *E. coli* under balanced feeding conditions, including N → K, G → S, W → C, F → C, and R → K misincorporations,[1] can also be explained by either a base change in the wobble position or a change involving $G^{mRNA}/U^{tRNA}$ or $U^{mRNA}/C^{tRNA}$ mismatches.

The favorable G → A base change requires a $G^{mRNA}/U^{tRNA}$ mismatch with G on the mRNA codon and U on the tRNA anticodon (Figure 1B). Logically, the $U^{mRNA}/G^{tRNA}$ mismatch involving the same two interacting bases should be similarly favorable. The $U^{mRNA}/G^{tRNA}$ mismatch should produce a U → C base change in the codons during translation. It is not clear why a U → C change is uncommon (M → T shown in Table 3 is the only observed example so far), but its infrequency is presumably related to the ribosomal proofreading process.[44] At the wobble position, however, while a $U^{mRNA}/G^{tRNA}$ mismatch is common, a $G^{mRNA}/U^{tRNA}$ mismatch is observed only when U on the tRNA is modified.[40] The common $G^{mRNA}/U^{tRNA}$ mismatch described here for the first and second bases is the opposite of that observed in the wobble position. The common $U^{mRNA}/G^{tRNA}$ mismatch at the wobble position, fortunately, does not produce any amino acid misincorporations because of the degeneracy of the genetic code. If the $G^{mRNA}/U^{tRNA}$ mismatch were common at the wobble position, we would observe an exceptionally high level of M → I misincorporation.

The Y → N (U → A base change) and M → T (U → C base change) misincorporations described here make an additional 10% [2 of 20 (Table 5)] of observed misincorporations.

**G/U Mismatches during Transcription.** The remaining unexplained misincorporations so far include P → L and P → S misincorporations in *E. coli*-expressed proteins, and N → S and D → G misincorporations in CHO-expressed proteins. The P → S misincorporations are observed in only one *E. coli*-expressed protein (Table 3), and the D → G misincorporation occurs at an extremely low level ($\sim 6 \times 10^{-6}$). N → S misincorporations are observed with a significant amount in only the clones derived from parent pool 4 of IgG2 mAb (X), indicating that it may be a special characteristic of that cell line (Table 1). Additionally, a higher level of N → S misincorporation is observed for cell culture condition B than for cell culture condition A.

Because mRNA synthesis is based on the antisense DNA as the template, the favorable G/U or G/T (T for thymine in DNA) mismatch[34] would produce a C → U or an A → G change (see Figure 1A). C → U and A → G changes are indeed observed frequently in mammals during mRNA synthesis[45−48] because of RNA editing by RNA-binding deaminases.[47,49,50] Interestingly, both the N → S and D → G misincorporations in CHO-expressed proteins not explained by the favorable mRNA/tRNA mismatches involve A → G base changes in their codons (Table 4), suggesting $G^{mRNA}/T^{DNA}$ mismatch or nonspecific activities of RNA editing enzymes as the possible causes for these misincorporations. In the *E. coli*-expressed proteins (Table 3), both the P → L and P → S misincorporations not explained by the favorable mRNA/tRNA mismatches involve a C → U base change, which can be explained by a $G^{DNA}/U^{mRNA}$ mismatch during transcription (Figure 1A). The C → U base changes in *E. coli* cannot be caused by RNA editing because of the lack of RNA editing functionality in prokaryotic cells.[47,50] The reason that the A → G change is not common in *E. coli*-expressed proteins (and was not observed in our study) is likely the lack of RNA editing functionality in *E. coli*, and also the possibility that the $G^{mRNA}/T^{DNA}$ mismatch required for the A → G change may not be as probable as the $G^{DNA}/U^{mRNA}$ mismatch required for the C → U change. G/U or G/T mismatches during transcription, or mRNA editing (C → U or A → G base changes), explain the remaining 20% [4 of 20 (Table 5)] of observed misincorporations. Note that the N → S misincorporation may also be caused by tRNA mischarging as described below.


dx.doi.org/10.1021/bi401002c | *Biochemistry* 2013, 52, 8165−8176

**tRNA Mischarging.** It has previously been concluded[19,20] that high levels of N → S misincorporation can be caused by mischarging of $tRNA^{Asn}$ with Ser under Asn starvation. The conclusion is supported in our data by significantly higher levels of N → S misincorporation in IgG2 mAb (X) produced under condition B than under condition A (Table 1), as the cell culture media contain a lower level of Asn under condition B than under condition A in later cell culture days (Figure S2 of the Supporting Information). Therefore, the N → S misincorporation observed in IgG2 mAb (X) may be caused by temporary Asn starvation.

**Factors Affecting Misincorporation.** It is widely believed that the error rate during protein translation depends on factors such as the type of organism, the genotype and phenotype of the transfected cell lines, the codon usage, and the cell culture conditions. Our data (Tables 1 and 2) indeed show that most misincorporations are related to both the cell line and the environment. For example, data in Table 1 indicate that the levels of N → K misincorporations are lower in clones 2-18 and 3-9 than in other clones, and N → S levels depend on both the cell line and cell culture conditions, with significant amount only in clones derived from pool 4 and under cell culture condition B [with reduced asparagine (Figure S2 of the Supporting Information)]. Data in Table 2 indicate that feeding condition B causes higher levels of N → K misincorporations, while feeding condition A causes higher levels of S → N and S → R misincorporations.

In this work, all the IgG2 samples were obtained from optimized cell cultures in which amino acids were generally not depleted. As a result, very low levels of misincorporations were observed (usually $<10^{-3}$). Although N → S misincorporation can be caused by temporary Asn starvation, Table 1 also clearly shows that the cell line used for the production has a greater impact on the levels of N → S misincorporations, indicating asparagine starvation is an overly simplified explanation for N → S misincorporations. For example, although feeding condition B2 for clone 3-9 has a very low level of Asn in the medium (Figure S2 of the Supporting Information), it produces very little N → S misincorporation, likely related to the rate of asparagine biosynthesis in that cell line. Upon comparison of levels of misincorporation in IgG2 mAb (X) to the amino acid concentration in the medium during cell culture, G → D, M → I, and V → I misincorporations show a negative correlation with concentrations of glycine, methionine, and valine, respectively, while other misincorporations show no rational correlation to the concentrations of corresponding amino acids (Table S2 of the Supporting Information). All these data indicate that both cell lines and feed conditions affect the levels of amino acid misincorporation.

**Hypothesis.** On the basis of the data presented in this report, we hypothesize that amino acid misincorporations are not random, with the most frequent misincorporations corresponding to either a G → A base change at any of the three codon positions or a C or U → A or G base change at the third codon position. These misincorporations are caused by G/U, C/U, or U/U mismatches during tRNA/mRNA codon recognition. Other misincorporations are more likely caused by transcriptional error when the misincorporated amino acid has a codon with a C → U or A → G difference, or tRNA mischarging, especially under nonoptimal feeding conditions and when the misincorporated amino acid has chemical properties similar to those of the intended amino acid.

Data suggest that most misincorporations under tested cell culture conditions (with balanced amino acids) are caused by mRNA/tRNA mismatching. It is believed that translation is a more error-prone step than transcription with error rates of $10^{-5}$ to $10^{-3}$, which match the error rates determined in this work. Only a small percentage of uncovered misincorporations were attributed to transcriptional errors and tRNA mischarging.

Note that the main cause of misincorporation may differ from system to system. For example, when a certain amino acid is depleted, the main cause of misincorporation may become mischarging of the tRNA with a wrong amino acid.

It is also important to point out that the amino acid misincorporations detected in recombinant proteins are those that survived the production and purification process. For example, in the case of CHO-expressed mAbs, the molecules containing amino acid substitutions must be successfully folded, assembled, secreted, and finally purified. If a substitution causes the molecule to fail any of these processes, it will not be detected. The level of misincorporation (and therefore the error rate) was determined in this work by the MS response of the corresponding peptides. Therefore, the determined levels might not be accurate when the ionization efficiency of the peptide is significantly changed after the substitution. The errors caused by the difference in ionization efficiencies can be as large as several-fold in extreme cases. The difference in determined levels of the same type of misincorporation at different locations is likely a combined effect of inaccuracies in the MS measurement, and different survival rates of the misincorporation at different locations during the production and purification process. Efforts have been made to map the error rates to the three-dimensional structure of the IgG molecule, as well as correlate the error rate to the solvent accessibility of each amino acid, and no solid conclusion has been made at this time, partly because of the lack of better statistics.

**Testing the Hypothesis.** The hypothesis proposed in this work states that under balanced nutrition conditions, most amino acid misincorporations are caused by a $G^{mRNA}/U^{tRNA}$ mismatch at any codon position, or C/U or U/U mismatches at the wobble position, during codon recognition. The hypothesis predicts that the levels of misincorporation depend strongly on the codons being used. Table 6 shows the predicted most possible amino acid misincorporations based on the hypothesis. Many of these misincorporations are codon-dependent (highlighted in Table 6). Amino acids with codon-dependent predictions are especially useful for testing the hypothesis. For example, according to the hypothesis, glycines encoded by GGG and GGA will generate G → E but not G → D misincorporation, while glycines encoded by GGU and GGC will generate G → D but not G → E misincorporation. To test the hypothesis, as part of a codon optimization effort,[17] the DNA sequence of an Fc fusion protein was varied to encode several amino acids by different codons, and the protein was expressed from *E. coli*. These amino acids included glycine, arginine, serine, and valine because they were observed in this work to be widely substituted by other amino acids. A total of 13 constructs were made with different DNA sequences, and 23 production runs were performed using these constructs (six constructs were expressed multiple times under different fermentation conditions).[17] Among the 23 runs, two runs generated significantly higher levels of misincorporation in many residues compared to the same construct expressed under a different fermentation condition, likely due to nonideal feeding conditions that caused misincorporation through different mechanisms (such as tRNA mischarging). The two abnormal runs were excluded from further data analyses. Table 7 shows the observed misincorpora-

**Table 6. Predicted Most Possible Amino Acid Misincorporations by a $G^{mRNA}/U^{tRNA}$ Mismatch or a Third-Base C/U or U/U Mismatch during Codon Recognition[a]**

| amino acid | codons | predicted misincorporated amino acids: G/U mismatch | predicted misincorporated amino acids: third-base mismatch |
|---|---|---|---|
| A | GCU, GCC, GCA, GCG | T | none |
| C | UGU, UGC | Y | W, stop |
| D | GAU, GAC | N | E |
| E | GAA, GAG | K | none |
| F | UUU, UUC | none | L |
| **G** | **GGA, GGG** | **E, R** | **none** |
| **G** | **GGU, GGC** | **D, S** | **none** |
| H | CAU, CAC | none | Q |
| **I** | **AUU, AUC** | **none** | **M** |
| **I** | **AUA** | **none** | **none** |
| K | AAA, AAG | none | none |
| L | all six codons | none | none |
| M | AUG | I | I |
| N | AAU, AAC | none | K |
| P | CCU, CCC, CCA, CCG | none | none |
| Q | CAA, CAG | none | none |
| **R** | **CGA, CGG** | **Q** | **none** |
| **R** | **CGU, CGC** | **H** | **none** |
| **R** | **AGA, AGG** | **K** | **none** |
| **S** | **AGU, AGC** | **N** | **R** |
| **S** | **UCU, UCC, UCA, UCG** | **none** | **none** |
| T | ACU, ACC, ACA, ACG | none | none |
| **V** | **GUU, GUC, GUA** | **I** | **none** |
| **V** | **GUG** | **M** | **none** |
| W | UGG | stop | stop |
| Y | UAU, UAC | none | stop |
| Stop | UAA, UAG, UGA | none | none |

[a]Amino acids with codon-dependent predictions are shown in bold.

**Table 7. Measured Levels of Misincorporation at G, R, S, and V Sites When Different Codons Were Used, Compared with the Predicted Levels (high or low) Based on the $G^{mRNA}/U^{tRNA}$ and Wobble Position Mismatch Hypothesis[a]**

| misincorporation | codon | prediction | average level of misincorporation ($\times 10^{-5}$) ± the standard deviation |
|---|---|---|---|
| G → D | **GGU** | **high** | **18 ± 36 (*n* = 26)** |
| | **GGC** | **high** | **325 ± 509 (*n* = 51)** |
| | GGA | low | 0 ± 0 (*n* = 6) |
| | GGG | low | 1 ± 0 (*n* = 2) |
| G → E | GGU | low | 3 ± 8 (*n* = 68) |
| | GGC | low | 2 ± 3 (*n* = 20) |
| | **GGA** | **high** | **269 ± 427 (*n* = 25)** |
| | **GGG** | **high** | **38 ± 42 (*n* = 26)** |
| G → R | GGU | low | 0 ± 1 (*n* = 28) |
| | GGC | low | 0 ± 0 (*n* = 12) |
| | **GGA** | **high** | **8 ± 12 (*n* = 16)** |
| | **GGG** | **high** | **1 ± 1 (*n* = 2)** |
| G → S | **GGU** | **high** | **7 ± 7 (*n* = 14)** |
| | **GGC** | **high** | **30 ± 25 (*n* = 10)** |
| | GGG | low | 2 ± 2 (*n* = 4) |
| R → K | CGU | low | 1 ± 1 (*n* = 9) |
| | CGC | low | 3 ± 2 (*n* = 2) |
| | **AGG** | **high** | **7 ± 6 (*n* = 5)** |
| R → Q | CGU | low | 0 ± 0 (*n* = 54) |
| | CGC | low | 0 ± 0 (*n* = 6) |
| | **CGA** | **high** | **29 ± 4 (*n* = 5)** |
| | **CGG** | **high** | **98 ± 76 (*n* = 16)** |
| S → N | UCU | low | 1 ± 1 (*n* = 15) |
| | UCC | low | 0 ± 1 (*n* = 3) |
| | UCG | low | 2 ± 1 (*n* = 2) |
| | **AGC** | **high** | **33 ± 19 (*n* = 71)** |
| V → I | **GUU** | **high** | **6 ± 4 (*n* = 31)** |
| | **GUC** | **high** | **34 ± 14 (*n* = 93)** |
| | **GUA** | **high** | **8 ± 3 (*n* = 2)** |
| | GUG | low | 2 ± 4 (*n* = 13) |
| | **GUG** | **high** | **9 ± 2 (*n* = 7)** |

[a]Observed misincorporation levels are represented as averages ± the standard deviation, and *n* is the number of codons with determined misincorporation levels.

tions at amino acid sites encoded by different codons, for the remaining 21 runs of the 13 constructs. Note that identified misincorporations were quantified across all runs. That is, whenever a misincorporation was identified in one run, its levels in all runs would be quantified, regardless of whether the misincorporation was present in that run. The observed levels of misincorporation agreed remarkably well with the predicted levels for the codons used. For example, when the codon of serine was changed from AGC to UCU, UCC, or UCG, the rate of S → N misincorporation changed from an average of $3.3 \times 10^{-4}$ to $\sim 1 \times 10^{-5}$. In fact, with only one exception (G → R for a glycine encoded by GGG, $\sim 1 \times 10^{-5}$), all misincorporations with high predicted levels had high observed levels (average levels of $6 \times 10^{-5}$ to $3 \times 10^{-3}$), and all misincorporations with low predicted levels had low observed levels (average levels of $< 5 \times 10^{-5}$). Of a large number of codon changes in these constructs, 210 of them involved a change from a codon matching the hypothesis to an alternative codon not matching the hypothesis. From these 210 codon changes, the corresponding amino acid misincorporations for 208 of them were either totally eliminated or greatly reduced (from an average level of $7.1 \times 10^{-4}$ to $1.1 \times 10^{-5}$), as predicted by the hypothesis. It is worthwhile to point out that when the level of a misincorporation was predicted to be high by the hypothesis, it was not necessarily high experimentally for all cell lines, as indicated by the large standard deviations listed in Table 7. This is presumably due to the availability of competing aminoacyl-tRNA's in the cells. However, the levels of misincorporation were usually very low when they were predicted to be low. Note that although the accuracy in the measured error rates is poor due to potential differences in ionization efficiencies, the precision and linearity of the measurement are usually good when the same species is quantified. Therefore, the relative difference in the rates of the same misincorporation in different runs is quite reliable.

**Amino Acid Misincorporations in Natural Proteins.** To answer the question of whether similar mechanisms of amino acid misincorporation occur in natural proteins, HSAs purified from three different individuals were digested with Lys-C and analyzed by LC−MS/MS peptide mapping, followed by MassAnalyzer data analysis to screen for amino acid misincorporations with single-base changes. The results are presented in Table 8. All 15 observed types of misincorporation (a total of 66 misincorporations observed) can be attributed to either $G^{mRNA}/U^{tRNA}$ (6 of 15) mismatches, wobble position U/U or C/U mismatches (2 of 15), $U^{mRNA}/G^{tRNA}$ (2 of 15) mismatches during translation, or $G^{DNA}/U^{mRNA}$ mismatches during transcription (5 of 15). These observations agree remarkably well with those


dx.doi.org/10.1021/bi401002c | *Biochemistry* 2013, 52, 8165−8176

**Table 8. Amino Acid Misincorporations Detected in Human Serum Albumin Purified from Three Human Subjects**

| misincorporation | codon[a] | no. of sites[b] | average level of misincorporation ($\times 10^{-5}$) subject 1 | subject 2 | subject 3 | base difference | possible cause transcription | translation |
|---|---|---|---|---|---|---|---|---|
| A → T | GCA | 3/17 | 6.4 | 5.6 | 6.9 | G1 → A | | G1/U mismatch |
| | GCG | 2/2 | 2.7 | 2.4 | 2.7 | | | |
| A → V | GCA | 3/17 | 3.3 | 3.7 | 4.2 | C2 → U | G/U mismatch or mRNA editing | |
| | GCC | 1/14 | 4.4 | 5.0 | 4.9 | | | |
| | GCU | 5/30 | 2.7 | 3.1 | 3.6 | | | |
| D → E | GAU | 2/25 | 2.8 | 2.1 | 2.2 | wobble | | U3/U or U3/C mismatch |
| D → N | GAU | 2/25 | 2.9 | 2.7 | 2.9 | G1 → A | | G1/U mismatch |
| E → K | GAA | 2/38 | 6.0 | 6.3 | 6.8 | G1 → A | | G1/U mismatch |
| | GAG | 4/24 | 1.4 | 2.3 | 2.2 | | | |
| F → L | UUC | 3/10 | 6.7 | 5.5 | 6.9 | wobble | | C3/U mismatch U3/U or U3/C mismatch |
| | UUU | 4/25 | 2.1 | 1.7 | 2.1 | | | |
| H → Y | CAU | 2/11 | 2.2 | 2.0 | 2.1 | C1 → U | G/U mismatch or mRNA editing | |
| L → F | CUU | 3/19 | 4.8 | 5.6 | 5.5 | C1 → U | G/U mismatch or mRNA editing | |
| M → I | AUG | 3/7 | 5.6 | 5.6 | 5.7 | wobble | | G3/U mismatch |
| M → T | AUG | 2/7 | 3.2 | 3.2 | 4.1 | U2 → C | | U2/G mismatch |
| R → K | AGA | 2/13 | 2.1 | 2.4 | 2.9 | G2 → A | | G2/U mismatch |
| | AGG | 1/5 | 2.4 | 2.5 | 3.4 | C2 → U | G/U mismatch or mRNA editing | |
| S → L | UCA | 2/6 | 6.0 | 6.1 | 6.1 | | | |
| | UCG | 1/3 | 7.4 | 9.3 | 10.3 | | | |
| T → I | ACA | 3/11 | 3.5 | 3.9 | 3.8 | C2 → U | G/U mismatch or mRNA editing | |
| | ACC | 3/9 | 2.8 | 3.2 | 3.5 | | | |
| | ACU | 1/7 | 2.1 | 2.0 | 1.9 | | | |
| V → A | GUA | 2/8 | 1.3 | 1.4 | 1.3 | U2 → C | | U2/G mismatch |
| | GUC | 1/7 | 2.9 | 1.7 | 2.4 | | | |
| | GUU | 2/12 | 2.8 | 2.3 | 2.9 | | | |
| V → I | GUA | 1/8 | 1.6 | 1.6 | 2.1 | G1 → A | | G1/U mismatch |
| | GUC | 1/7 | 3.9 | 3.0 | 4.3 | | | |
| | GUU | 5/12 | 4.8 | 4.2 | 4.9 | | | |

[a]DNA sequence obtained from NCBI GenBank. [b]Number of observed amino acid misincorporations compared to the total number of specified codons for HSA.

made with overexpressed recombinant proteins, indicating the validity of the hypothesis for natural proteins. Misincorporations observed in natural HSA were generally lower than $10^{-4}$, and no significant differences were observed in the level of these misincorporations among different individuals.

## CONCLUSION

Results presented in this work demonstrate that G/U mismatches during translation and transcription and C/U and/or U/U mismatches at the wobble position during translation are key causes of amino acid misincorporation for recombinant as well natural proteins under balanced nutrient conditions. This knowledge can guide the codon optimization process in the biopharmaceutical industry to reduce the level of misincorporation during protein engineering and minimize the number of tested constructs. It will help to recognize misincorporations caused by starvation and optimize feeding regiment. The developed methodology can be used for screening DNA mutations during early stages of product development, detecting low levels of misincorporations for the support of codon selection, feed optimization, and improving our general understanding of the cell machinery during protein synthesis.

## ASSOCIATED CONTENT

### Supporting Information

Discussions regarding experimental considerations and performance aspects of the methodology, common modifications that can be misidentified as amino acid substitutions, correlation of misincorporation with the amino acid concentration in the cell culture medium, an example of the peptide identification process by MassAnalyzer, and asparagine concentrations in the cell culture media of different cell lines that produced different levels of N → S misincorporation. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

### Corresponding Author

*Process and Product Development, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320. E-mail: zzhang@amgen.com. Telephone: (805) 447-7783. Fax: (805) 376-2354.

### Funding

Work supported by Amgen Inc.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Mark Leo Michaels, Ed Belouski, Mark Berge, Hedieh Barkhordarian, Frank Abroson, Dwight Winters, and Steve Smith for making the constructs and expressing the proteins for testing the hypothesis. We also thank Greg Flynn, James "Rusty" Lipford, Rohini Deshpande, and Li Zhang for helpful discussions and Li Zhang and Eugene Babcock for providing cell culture amino acid data.

## ■ ABBREVIATIONS

MS/MS, tandem mass spectrometry; LC−MS/MS, liquid chromatography online with tandem mass spectrometry; Fc, fragment crystallizable region of an antibody; LC, liquid chromatography or light chain; HC, heavy chain; PIE, precursor ion exclusion; HSA, human serum albumin.

## ■ REFERENCES

(1) Parker, J. (1989) Errors and alternatives in reading the universal genetic code. *Microbiol. Mol. Biol. Rev. 53*, 273−298.
(2) Drummond, D. A., and Wilke, C. O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet. 10*, 715−724.
(3) Rodnina, M. V., and Wintermeyer, W. (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: Kinetic and structural mechanisms. *Annu. Rev. Biochem. 70*, 415−435.
(4) Francklyn, C. S. (2008) DNA polymerases and aminoacyl-tRNA synthetases: Shared mechanisms for ensuring the fidelity of gene expression. *Biochemistry 47*, 11695−11703.
(5) Zaher, H. S., and Green, R. (2009) Fidelity at the molecular level: Lessons from protein synthesis. *Cell 136*, 746−762.
(6) Ling, J., Reynolds, N., and Ibba, M. (2009) Aminoacyl-tRNA synthesis and translational quality control. *Annu. Rev. Microbiol. 63*, 61−78.
(7) Reynolds, N. M., Lazazzera, B. A., and Ibba, M. (2010) Cellular mechanisms that control mistranslation. *Nat. Rev. Microbiol. 8*, 849−856.
(8) Edelmann, P., and Gallant, J. (1977) Mistranslation in *E. coli*. *Cell 10*, 131−137.
(9) Parker, J., and Friesen, J. D. (1980) "Two out of three" codon reading leading to mistranslation in vivo. *Mol. Gen. Genet. 177*, 439−445.
(10) Ellis, N., and Gallant, J. (1982) An estimate of the global error frequency in translation. *Mol. Gen. Genet. 188*, 169−172.
(11) Bouadloun, F., Donner, D., and Kurland, C. (1983) Codon-specific missense errors in vivo. *EMBO J. 2*, 1351−1356.
(12) Toth, M. J., Murgola, E. J., and Schimmel, P. (1988) Evidence for a unique first position codon-anticodon mismatch in vivo. *J. Mol. Biol. 201*, 451−454.
(13) Ogle, J. M., and Ramakrishnan, V. (2005) Structural insights into translational fidelity. *Annu. Rev. Biochem. 74*, 129−177.
(14) Kramer, E. B., and Farabaugh, P. J. (2007) The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA 13*, 87−96.
(15) 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 *E. coli*. *J. Mol. Biol. 262*, 407−412.
(16) Yu, X. C., Borisov, O. V., Alvarez, M., Michels, D. A., Wang, Y. J., and Ling, V. (2009) Identification of codon-specific serine to asparagine mistranslation in recombinant monoclonal antibodies by high-resolution mass spectrometry. *Anal. Chem. 81*, 9282−9290.
(17) Hutterer, K. M., Zhang, Z., Michaels, M. L., Belouski, E., Hong, R. W., Shah, B., Berge, M., Barkhordarian, H., Le, E., Smith, S., Winters, D., Abroson, F., Hecht, R., and Liu, J. (2012) Targeted codon optimization improves translational fidelity for an Fc fusion protein. *Biotechnol. Bioeng. 109*, 2770−2777.
(18) Parker, J., Pollard, J. W., Friesen, J. D., and Stanners, C. P. (1978) Stuttering: High-level mistranslation in animal and bacterial cells. *Proc. Natl. Acad. Sci. U.S.A. 75*, 1091−1095.
(19) Wen, D., Vecchi, M. M., Gu, S., Su, L., Dolnikova, J., Huang, Y. M., Foley, S. F., Garber, E., Pederson, N., and Meier, W. (2009) Discovery and investigation of misincorporation of serine at asparagine positions in recombinant proteins expressed in Chinese hamster ovary cells. *J. Biol. Chem. 284*, 32686−32694.
(20) Khetan, A., Huang, Y., Dolnikova, J., Pederson, N. E., Wen, D., Yusuf Makagiansar, H., Chen, P., and Ryll, T. (2010) Control of misincorporation of serine for asparagine during antibody production using CHO cells. *Biotechnol. Bioeng. 107*, 116−123.
(21) Zhang, Z. (2009) Large-scale identification and quantification of covalent modifications in therapeutic proteins. *Anal. Chem. 81*, 8354−8364.
(22) Ren, D., Pipes, G. D., Liu, D., Shih, L.-Y., Nichols, A. C., Treuheit, M. J., Brems, D. N., and Bondarenko, P. V. (2009) An improved trypsin digestion method minimizes digestion-induced modifications on proteins. *Anal. Biochem. 392*, 12−21.
(23) Zhang, Z. (2012) Automated precursor ion exclusion during LC-MS/MS data acquisition for optimal ion identification. *J. Am. Soc. Mass Spectrom. 23*, 1400−1407.
(24) Zhang, Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem. 76*, 3908−3922.
(25) Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem. 77*, 6364−6373.
(26) Zhang, Z., and Shah, B. (2010) Prediction of collision-induced dissociation spectra of common N-glycopeptides for glycoform identification. *Anal. Chem. 82*, 10194−10202.
(27) Zhang, Z. (2011) Prediction of collision-induced dissociation spectra of peptides with post-translational or process-induced modifications. *Anal. Chem. 83*, 8642−8651.
(28) Zhang, Z. (2012) Retention time alignment of LC/MS data by a divide-and-conquer algorithm. *J. Am. Soc. Mass Spectrom. 23*, 764−772.
(29) Buxton, G. V., Greenstock, C. L., Helman, W. P., and Ross, A. B. (1988) Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals. *J. Phys. Chem. Ref. Data 17*, 513−886.
(30) Xu, G., and Chance, M. R. (2005) Radiolytic modification and reactivity of amino acid residues serving as structural probes for protein footprinting. *Anal. Chem. 77*, 4549−4555.
(31) Tomita, M., Irie, M., and Ukita, T. (1969) Sensitized photooxidation of histidine and its derivatives. Products and mechanism of the reaction. *Biochemistry 8*, 5149−5160.
(32) Schaefer, H., Chamrad, D. C., Marcus, K., Reidegeld, K. A., Bluggel, M., and Meyer, H. E. (2005) Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. *Proteomics 5*, 846−852.
(33) Fodor, S., and Zhang, Z. (2006) Rearrangement of terminal amino acid residues in peptides by protease-catalyzed intramolecular transpeptidation. *Anal. Biochem. 356*, 282−290.
(34) Müller, U. R., and Fitch, W. M. (1985) The biological significance of G-T/G-U mispairing in nucleic acid secondary structure. *J. Theor. Biol. 117*, 119−126.
(35) Sugimoto, N., Kierzek, R., Freier, S. M., and Turner, D. H. (1986) Energetics of internal GU mismatches in ribooligonucleotide helixes. *Biochemistry 25*, 5755−5759.
(36) Limmer, S. (1997) Mismatch base pairs in RNA. *Prog. Nucleic Acid Res. Mol. Biol. 57*, 1−39.
(37) Varani, G., and McClain, W. H. (2000) The G·U wobble base pair. *EMBO Rep. 1*, 18−23.
(38) Uhlenbeck, O. C., Baller, J., and Doty, P. (1970) Complementary oligonucleotide binding to the anticodon loop of fMet-transfer RNA. *Nature 225*, 508−510.
(39) Freier, S. M., Kierzek, R., Caruthers, M. H., Neilson, T., and Turner, D. H. (1986) Free energy contributions of G-U and other terminal mismatches to helix stability. *Biochemistry 25*, 3209−3213.
(40) Agris, P. F., Vendeix, F. A. P., and Graham, W. D. (2007) tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol. 366*, 1−13.

(41) Loftfield, R. (1963) The frequency of errors in protein biosynthesis. *Biochem. J. 89*, 82−92.
(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* al homeodomain expressed in *Escherichia coli*. *Protein Sci. 7*, 500−503.
(43) McNulty, D. E., Claffee, B. A., Huddleston, M. J., and Kane, J. F. (2003) Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Expression Purif. 27*, 365−374.
(44) Ogle, J. M., Murphy, F. V., IV, Tarry, M. J., and Ramakrishnan, V. (2002) Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell 111*, 721−732.
(45) Chen, S.-H., Habib, G., Yang, C.-Y., Gu, Z.-W., Lee, B. R., Weng, S.-A., Silberman, S. R., Cai, S.-J., Deslypere, J., and Rosseneu, M. (1987) Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science 238*, 363−366.
(46) Powell, L. M., Wallis, S. C., Pease, R. J., Edwards, Y. H., Knott, T. J., and Scott, J. (1987) A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell 50*, 831−840.
(47) Li, J. B., Levanon, E. Y., Yoon, J.-K., Aach, J., Xie, B., LeProust, E., Zhang, K., Gao, Y., and Church, G. M. (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science 324*, 1210−1213.
(48) Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science 333*, 53−58.
(49) Chester, A., Scott, J., Anant, S., and Navaratnam, N. (2000) RNA editing: Cytidine to uridine conversion in apolipoprotein B mRNA. *Biochim. Biophys. Acta 1494*, 1−13.
(50) Bass, B. L. (2002) RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem. 71*, 817−846.