

# A Bernoulli Process Approximation for the Reverse Translation of Protein to DNA/mRNA

<sup>1</sup>Syed Ahsan, <sup>2</sup>Amjad Farooq, <sup>3</sup>Abad Shah

<sup>1,3</sup> Al-khawarizmi Institute of Computer Science

<sup>2</sup>Department of Computer Science, University of Engineering and Technology, Lahore  
ahsancs@hotmail.com

**Abstract:** In proteomics, to find the genomic localization of a gene, a protein may be interpreted back to its DNA/mRNA using the method of reverse translation. A particular amino acid may be translated into more than one codon as the uncertainties exist in the reverse translation of a protein. In this paper, we propose a Bernoulli process approximation based on the usage of frequency distribution for the reverse translation of an amino acid to its DNA/mRNA. A comparison of our proposed method with the existing conventional procedure that is based on random selection of codon, is also presented. Preliminary results of the proposed process are encouraging and it shows improved accuracy and consistency. [Journal of American Science 2010; 6(9):60-64]. (ISSN: 1545-1003).

**Keywords:** Bioinformatics, Proteomics, Reverse Translation of mRNA, genetic code, Bernoulli process, Frequency distribution, Binomial probability

## 1. Introduction

To get a balanced production of proteins required for the endurance of the cell transcription to mRNA and a consequent translation to proteins take place for all living cells containing DNA. The total complement of proteins present in a cell or cell type is known as its proteome, and the study of such large-scale data sets defines the field of proteomics. In proteomics, interpreting protein back to its DNA/mRNA is of immense significance because to find the genomic localization of the gene, a protein may be interpreted back to its DNA/mRNA using the method of reverse translation. A specific protein render from the translation of a DNA sequence while the reverse translation may not map exactly to a specific protein in accordance with the genetic code because a particular amino acid can be translated into more than one codon. This creates uncertainties in the reverse translation of a particular protein that is referred to as the problem of uncertainty of reverse translation (Posele et al., 19898).

The problem of uncertainty in the reverse translation is due to the degeneracy of the genetic code, and the diversity and range available for the codon usage in the translation of a protein back to its DNA/mRNA and vice versa. Various approaches are reported in the literature to resolve the uncertainties of the reverse translation. One of such approach is to *select a codon in random fashion*. However, this *hit-and-trial* approach is vulnerable to inaccuracy and inconsistency as picking a codon randomly might not always result in choosing the best codon. There is

another approach, called the *selection of a codon in accordance with the frequency* by which it has occurred in a specific organism (Jagdes, 2004; Grote et al., 2005). Choosing the codon from a distribution of premeditated codon the usage of frequencies increases the possibility of choosing correct codons for the DNA/mRNA sequence Grote et al., 2005).

In this paper, we propose a Bernoulli process approximation for the translation of the amino acid residue back to DNA/mRNA based on the codon usage of the frequency distribution in a specific organism. Through a statistical illustration, we determine that the Bernoulli process approximation based on the frequency distribution based codon results in a more accurate reverse translation as compared to the conventional procedure based on random selections of codons.

The remainder of the paper is divided into four sections. Section 2 covers the related work. In Section 4, we propose the Bernoulli process approximation for reverse translation. Section 4 contains experimental results and also the advantages of our proposed technique are discussed.

## 2. Related Work

Codons are triplets of nucleotides that together specify an amino acid residue in a polypeptide chain. Most organisms use 20 or 21 amino acids to make their polypeptides, which are proteins or protein precursors. Because there are four possible

nucleotides, adenine (A), guanine (G), cytosine (C) and thymine (T) in DNA, there are 64 possible triplets to recognize only 20 amino acids plus the translation termination signal (Posele et al., 1988). Because of this redundancy, all but two amino acids are coded for by more than one triplet. Different organisms often show particular *preferences* for one of the several codons that encode the same given amino acid (Plotkin et al., 2006).

It is generally acknowledged that codon preferences reflect a balance between mutational biases and natural selection for translational optimization. Optimal codons in fast-growing microorganisms, like *Escherichia coli*, reflect the composition of their respective genomic tRNA pool (Posele et al., 1988). It is thought that optimal codons help to achieve faster translation rates and high accuracy. As a result of these factors, translational selection is expected to be stronger in highly expressed genes, as is indeed the case for the above-mentioned organisms (Plotkin et al., 2006).

A codon usage frequency table can assist in the assessment of the codon preference in the various species. The significance of taking into account the disparity in an amino acid codon usage comes from the degenerate nature of the genetic code (Lambrix et al., 2009). Most of the twenty-one (21) amino acids are encoded by one or more codon/triplets (Yury et al., 2008). This encoding is of immense significance in the characterization and functional examination of coding regions of a specific gene or a group of genes. Genes are present in an organism or in related species, usually demonstrate the same pattern of codon usage (Posele et al., 1988). Assessments of codon usage within and between various species can disclose the degree of pattern conservation and allow the examination whether the conservation is across all the members of the same specie or just across the members of a common phylogenetic family of an individual specie (Ahsan and Shah, 2006; Jagadesh, 2004). Furthermore, this information is precious in devising the primers for polymerase chain reaction (PCR) and assessment of exon precincts (Plotkin et al., 2006).

Triplet	Amino acid	Fraction	Frequency/Thousand
TTT F 0.58 22.1 ( 80995) TTC F 0.42 16.0 ( 58774) TTA L 0.14 14.3 (52382) TTG L 0.13 13.0 (47500)	TCT S 0.17 10.4 (38027) TCC S 0.15 9.1 (33430) TCA S 0.14 8.9 (32715) TCG S 0.14 8.5 (31146)	TAT Y 0.59 17.5 ( 63937) TAC Y 0.41 12.2 ( 44631) TAA * 0.61 2.0 ( 7356) TAG * 0.09 0.3 ( 989)	TGT C 0.46 5.2 (19138) TGC C 0.54 6.1 (22188) TGA * 0.30 1.0 (3623) TGG W 1.00 13.9 (50991)
CTT L 0.12 11.9 ( 43449) CTC L 0.10 10.2 ( 37347) CTA L 0.04 4.2 ( 15409) CTG L 0.47 48.4 ( 177210)	CCT P 0.18 7.5 ( 27340) CCC P 0.13 5.4 ( 19666) CCA P 0.20 8.6 ( 31534) CCG P 0.49 20.9 (76644)	CAT H 0.57 12.5 (45879) CAC H 0.43 9.3 (34078) CAA Q 0.34 14.6 (53394)	CGT R 0.36 20.0 (73197) CGC R 0.36 19.7 (72212) CGA R 0.07 3.8 (13844) CGG R 0.11 5.9 (21552)
ATT I 0.49 29.8 (109072) ATC I 0.39 23.7 (86796) ATA I 0.11 6.8 (24984) ATG I 1.00 26.4 (96695)	ACT T 0.19 10.3 (37842) ACC T 0.40 22.0 (80547) ACA T 0.17 9.3 (33910) ACG T 0.25 13.7 (50269)	AAT N 0.49 20.6 (75436) AAC N 0.51 21.4 (78443) AAA K 0.74 35.3 (129137) AAG K 0.26 12.4 (45459)	AGT S 0.16 9.9 (36097) AGC S 0.25 15.2 (55551) AGA R 0.07 3.6 (13152) AGG R 0.04 2.1 (7607)
GTT V 0.28 19.8 (72584) GTC V 0.20 14.3 (52439) GTA V 0.17 11.6 (42420) GTG V 0.35 24.4 (89265)	GCT A 0.18 17.1 (62479) GCC A 0.26 24.2 (88721) GCA A 0.23 21.2 (77547) GCG A 0.33 30.1 (110308)	GAT D 0.63 32.7 (119939) GAC D 0.37 19.2 (70394) GAA E 0.68 39.1 (143353) GAG E 0.32 18.7 (68609)	GGT G 0.35 25.5 (93325) GGC G 0.37 27.1 (99390) GGA G 0.13 9.5 (34799) GGG G 0.15 11.3 (41277)

Table 1: *Escherichia coli* (Genetic Code: Standard)

### 3. Bernoulli Process Approximation for the reverse translation

Protein sequences can be abstractly modeled as residues of amino acids connected together linearly (Jagdeh, 2004). A specific protein sequence is an alphabet that is comprised over twenty letters symbol that is translated from selected stretches of a particular DNA. A predefined translation table is used to translate each 3 letters of DNA to one amino-acid (AA) [see table1]. We can approximate the selection of a codon for the reverse translation to DNA/mRNA through Bernoulli process. The accuracy of this approximation is relative to the maximum number of codons reverse translated in the original DNA/mRNA sequence prior to actual protein translation. The obtained DNA sequence may not be an exact match to the original sequence due to the degeneracy of the genetic code.

Let us assume that the codons in a protein sequence have already been calculated and we have the codon usage frequency data. All the 64 possible codons are categorized along with the categories of 21 amino acids based on their behavioral properties. From each category of amino acid only one codon, at a time, can participate to code that amino acid. Let

$N$  = total number of amino acids are involved to form a protein sequence.

$K$  = number of codons that can be encoded by a particular amino acid.

$x$  = current position at which an amino acid is being reverse translated into a codon.

$\xi$  = codon selected for the reverse translation of a particular amino acid at position  $x$ .

$\mu_x$  = probability of the occurrence of a specific codon at position  $x$ .

Then the Bernoulli process approximation for the reverse translation can be formalized in the terms of probability space  $PS$  is (Gordon, 1997)

$$PS = ( \Omega, \mathcal{F} ) : P( \cdot ) = 1 \quad (1)$$

In equation (1)  $\Omega$  is the set of all possible codon that one amino acid can code

Let  $\mu$  be a random variable over the set  $\{0, 1\}$ , so that for every possible codon  $\omega$  at specific location  $x$ , then

$$\mu_x(\omega) = 1 \text{ with probability } \mu_x \quad (2)$$

$$\text{and } \mu_x(\omega) = 0 \text{ with probability } 1 - \mu_x \quad (3)$$

In equation (2) 1 signifies the 'success' of selection of a codon for translation. In equation (3), 0 signifies the 'failure' of selection of a codon for translation.

Formally  $\mu$  is

$$\mu_x(\omega) = \begin{cases} 1 & \omega = \xi \\ 0 & \omega = \text{any codon other than } \xi \end{cases}$$

Such that for every random variable  $\mu$  (*e. amino acid*) at any specific location  $x \geq 1$  in the protein sequence with its possible set of codons  $\Omega$  of total number  $K$ , there exists a family of random variables  $\{\mu_x | \mu_x(\cdot) \text{ for } x \geq 1\}$  (5)

Such that for a distinct value of  $x = \{1, 2, 3, \dots, N\}$ , one gets a distinct random variable that constitutes the reverse translation process into a stochastic process a Bernoulli process [12].

Now consider the reverse translation as a Bernoulli process with  $\{\mu_x | x = 1, 2, 3, 4, \dots, N\}$ . Let be the location of amino acids (starting from  $x = 1, 2, 3, 4, \dots, N$ ) at which that an amino acid decodes the codon  $\xi$  for reverse translation i.e.

$$= x \text{ if } \mu_1 = \mu_2 = \mu_3 = \dots = \mu_{x-1} = 0 \text{ and } \mu_x = 1 \quad (6)$$

Then  $\mu_x$  is *geometric probability mass function (pmf)* with parameter  $\mu_x$  i.e.

Now let  $n$  be the consecutive number of amino acids for which codon  $\xi$  codon has been used for the reverse translation in the  $\mu_x = x = 1$  to  $n$

### 4. Experimental Results and Discussion

In this section we compare the results of reverse translations obtained through the *Bernoulli process approximation* with the conventional *random based selection method*. We have used *Escherichia coli*, which is one of the main species of bacteria living in the lower intestines of mammals. For the reverse translation achieved through *Bernoulli process approximation*, we have used the *codon usage frequency distribution* for the standard genetic code of *Escherichia coli*. We performed our experiment on the protein of *Aldehyde-alcohol Aldehyde dehydrogenase (adhE)* and obtained the reverse translation into the mRNA of this protein with the Bernoulli approximation defined through the *codon usage frequency distribution* of the whole *Escherichia coli* class and compared it with translation archived through the *random-codon selection*.

*Alcohol dehydrogenases* are a group of *dehydrogenase enzymes* that occur in many organisms and facilitate the interconversion between *alcohols* and *aldehydes* or *ketones*. In humans and many other animals, they serve to break down alcohols which could otherwise be toxic; in yeast and many bacteria, some alcohol dehydrogenases

catalyze the opposite reaction as part of fermentation. humans, the enzyme is contained in the lining of the stomach and in the liver. It catalyzes the oxidation of ethanol to acetaldehyde:

We selected *Alanine* amino acids from the whole protein sequence of *adhE* and specifically studied the behaviors of their reverse translation into the codons they generated. The total length of the *adhE* protein sequence was 885 and *Alanine* was found on 159 locations in the whole protein sequence. A *lanine* can be reverse translated into four different codons which include *GCU*, *GCC*, *GCA* or *GCG*. By means of picking either one by random selection we will get an *Alanine*. According to the *codon usage frequency distribution* [3], the most frequent codon, coding for an *Alanine* in *Escherichia coli* is *GCG*, encoding approximately 33.4% of all *Alanine*. Codon *GCC* codes about 26.8%, *GCA* about 22.9% and lastly *GCU* approximately 18.3%. We studied the behaviors of reverse translation of the above 4 possible codons on the 159 possible location where *Alanine* occurred in the *Aldehyde-alcohol Aehydrogenase (adhE)* protein sequence and then compared it with the actual mRNA of *Aldehyde-alcohol Aehydrogenase (adhE)*.

In the Bernoulli process approximation, the probability space  $(\Omega, \mathcal{F}, P) = \{(GCU, 0.18), (GCA, 0.23), (GCC, 0.26), (GCG, 0.33)\}$  for each possible codon for *Alanine*. Either one of the codon can code for *Alanine* at a time so when one codon codes it has the probability as 1 denoted by  $\epsilon$  and rest of the 3 codons have probability as 0. In this experiment we checked for each possible codons by assigning each one the probability one in 4 different cases and compared them with actual mRNA translation afterwards. By calculating the *geometrical* and *binomial* probability functions for locations where *Alanine* existed in the protein residue, we computed the probability of each codon and obtained reverse translations accordingly. As the probability assigned to each codon is estimated from the *codon usage frequency distribution* respective measures of geometrical and binomial mass functions reflected the behavior of *codon usage frequency distribution*.

In case of Bernoulli approximated reverse translation, 29 *GCA* codons were accurately reverse translated against 35 in the original mRNA translation and 46 *GCC*, 63 *GCG*, 21 *GCU* were accurately reverse translated against 52 *GCC*, 55 *GCG*, 17 *GCU* codons of *Alanine* respectively.

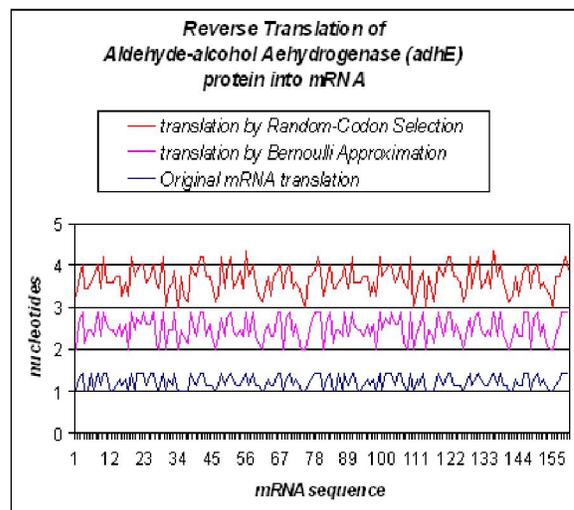
In case of random-selected codon reverse translation, 49 *GCA* codons were accurately reverse translated against 35 in the original mRNA translation and 31 *GCC*, 44 *GCG*, 35 *GCU* were accurately reverse translated against 52 *GCC*, 55 *GCG*, 17 *GCU* codons of *Alanine* respectively. The

comparison provides the evidence that the reverse translation based on *codon usage frequency distribution* with the relevant probability assigned is far more accurate than the

Codon	Original mRNA Translation	Bernoulli approximation reverse translation	Random selection
GCA	35	29	49
GCC	52	46	31
GCG	55	63	44
GCU	17	21	35

**Table 2:** Bernoulli approximation using frequency distribution maps better against the original mRNA translation

*random-selection based reverse translation* Comparison of achieved mRNA translations in Figure 3 shows that *Bernoulli approximation* based reverse translation maps more accurately to the original mRNA translations as compared to *random-selection* based reverse translation. In Figure 4(a) the graph of the *Alanine* amino acid also shows that the distribution of each codon in case of *Bernoulli approximation* based reverse translation is far more closer to original mRNA translation as compared to *random-selection* based reverse translation.



**Figure 1 :** Comparison of the three reverse translations obtained for *Aldehyde-alcohol Aehydrogenase (adhE)*

In Figure 4(a) the graph of the distribution of *Alanine* amino acid in the whole reverse translations, shows that the distribution of overall

Alanine, in case of Bernoulli approximation based reverse translation is closer to original mRNA translation than the random-selection based reverse translation.

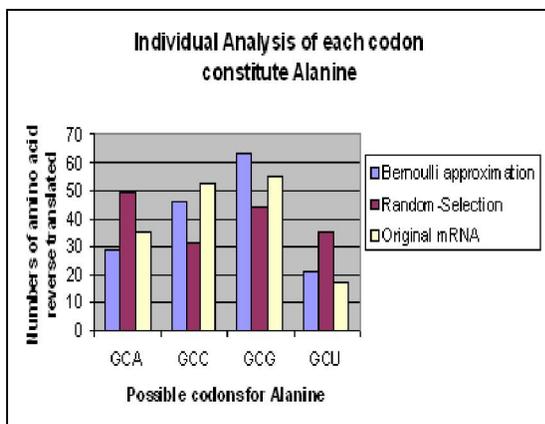
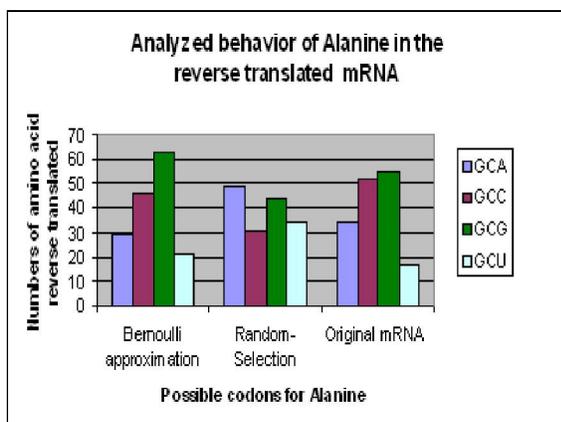


Figure 24(a): Analysis of reverse translations for Alanine



(b): Overall pattern of Alanine in reverse translations obtained

## 5. Conclusions

In this paper we discussed the importance of reverse translation in proteomics and its accompanying problem of uncertainty. This uncertainty arises as different organisms often show particular preferences for one of the several codons that encode the same given amino acid. For a more accurate and efficient mapping in reverse translation, we feel that codon selection based on codon usage frequency can give better results as compared to random selection of codon. We proposed and used Bernoulli process approximation for the translation of the amino acid residue back to DNA/mRNA based on the codon usage of the frequency distribution in a specific organism. Through a statistical illustration, we determined that the Bernoulli process approximation

based on the frequency distribution based codon results in a more accurate reverse translation as compared to the conventional procedure based on random selections of codons.

## Acknowledgement

This research work has been supported by the "Higher Education Commission of Pakistan", and the University of Engineering and Technology, Lahore.

## References

1. Ahsan S., Shah A. "Biological Databanks: Distribution, Heterogeneity, Diversity and Provenance", 7<sup>th</sup> Workshop on Distributed Data and Structures (WDAS 2006).
2. Amamath Gupta San Diego Supercomputer Center, University Of California San Diego La Jolla, CA 92130 "Life Science Research and Data Management What can they give each other?"
3. Gordon H. Discrete Probability. Springer, 1997.
4. Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC, Jahn D J-Cat: A novel tool for adopt codon usage of a target gene to its potential expression host. Nucleic Acids Res 2005, 33 (Web Server Issue):W 526-31
5. Jagadesh H. "Data Management For Life Sciences Research". SIGMOD RECORD, June 2004.
6. Jagdish H.V. "Data Management For Life Sciences, Report Of The NSF/NLM workshop on Data Management for Molecular and Cell Biology" National Library For Medicine Nov4, 2004.
7. Karl R. Popper. The Logic of Scientific Discovery
8. Lambrix P., Strömback L., Tan H. Information Integration in Bioinformatics with Ontologies and Standards. REVERSE 2009: 343-376
9. Plotkin J., Dushoff J., Desai M., and Fraser H. Codon Usage and Selection on Proteins. Journal of Molecular Evolution. Springer, 2006.
10. Posele G, Attimonelli M, Lioni S. A back translation method based on codon usage strategy. Nucleic Acids Res. 1988, 16(5):1715-28
11. Yury V., Moyer D., Rob M., Peter C., Thodoros T., Thierry B., Theo G, Henry D., Ian I., Jaek R., Nancy F. Design and Analysis of Quantitative Differential Proteomics Investigations Using LC-MS Technology. J. Bioinformatics and Computational Biology 6(1): 107-123 (2008).

5/6/2010