Neural Network Analysis of Protein Synthesis Termination Signal Efficiency

Michael Watts[a], Louise Major[b], Nik Kasabov[a], Warren Tate[b]

[a]Department of Information Science — [b]Department of Biochemistry
University of Otago, PO Box 56,
Dunedin, New Zealand
E-mail: mike@kel.otago.ac.nz

Abstract

The application of multi-layer perceptrons to the protein synthesis termination signal efficiency problem from molecular biology is presented. It is shown that by observing and analysing the performance of ANNs over this problem, useful insights into the biological processes involved can be gained.

1 Introduction

The protein synthesis termination signal efficiency problem can be stated as follows: given a specific termination codon in a messenger RNA strand, how is the efficiency of the termination codon modified by the three bases that follow it?

The answers to that question can be determined experimentally, and the biological background of this problem is described in Sections 2 and 3. The biological mechanisms involved, however, are somewhat harder to determine. It is in an attempt to determine the nature of these mechanisms that artificial neural networks (ANN) may be applied to this problem. This paper describes a preliminary study of this application.

The first question that arises, is can an ANN even model this problem? That question is addressed in Section 5, along with questions about the scheme by which RNA bases are represented within the network. This is followed by an investigation of the importance of various bases to one specific stop codon, the UGA codon, in Section 6. Further comparisons between the three codons investigated are presented in Section 7. Some directions for future research are discussed in Section 8, and the conclusions drawn from the research are presented in Section 9.

2 Biological background

Genetic information is stored in the nucleotide, or base, sequence of DNA. There are four DNA bases, adenine (A), thymine (T), cytosine (C) and guanine(G). In the double stranded DNA molecule A and T form base pairs, as do C and G. When the bases of two regions of nucleic acid have sequences which can form base pairs they are described as being complementary. From DNA a complementary strand of RNA can be transcribed, containing the RNA bases A, C, G and uracil (U). The mRNA are then translated into proteins.

Proteins are chains of amino acids linked by peptide bonds. Protein synthesis occurs on highly specialised macromolecules called ribosomes. The translation of mRNA into proteins requires recognition of triplets of mRNA bases (called codons) by complementary sequences in the anticodon loop of tRNA molecules which have specific amino acids attached to one end of their RNA. When two tRNA molecules have bound to adjacent mRNA codons on the ribosome the amino acid from one tRNA is transferred (covalently bound) to the amino acid on the other tRNA in a peptidyl transferase reaction catalysed by the ribosome. The protein is elongated in an iterative process.

With the four different bases in mRNA being read as triplet codons there are 4^3, or 64 possible triplet combinations. In the redundant genetic code sixty-one of the triplet codons code for twenty amino acids, the remaining three codons (UAA, UAG and UGA) are termination or stop codons which signal the end of protein synthesis from that mRNA. The specificity of the genetic code comes from the specific recognition of codons by the tRNA molecules, and the tRNA molecules are each charged with a specific amino acid.

The three termination codons are recognised by decoding release factors. The decoding release factors are proteins which cause the ribosome to catalyse peptidyl hydrolysis instead of peptidyl transferase, hence releasing the newly synthesised protein from the ribosome. In *Escherichia coli* (*E. coli*, the model organism used in this study) there are two decoding release factors, RF1 and RF2. RF1 terminates protein synthesis at UAA and UAG stop codons and RF2 terminates protein synthesis at UAA and UGA stop codons.

3 Effect of stop codon context on termination efficiency

The sequence immediately after stop codons affects how efficiently the stop codon is decoded by release factors [3, 1]. The efficiency of termination of protein synthesis, at stop codons with different downstream sequences, was measured *in vivo* in *E. coli*. In the experimental system termination was measured in competition with frameshifting (where instead of reading triplet code as 123 123 123, one base is missed out, so the same sequence is read as 123 231 23, which will encode different amino acids). The likelihood of frameshifting occurring depends on the speed with which the termination codon is decoded. If the termination signal is 'strong' very little frameshifting product is produced, while if the stop signal is 'slow', very little termination product is produced. Termination produced a shorter protein than frameshifting. The amount of each of these two proteins could be measured, and the termination percentage efficiency was calculated as per Equation 1:

$$\% = \frac{termination\ product}{termination + frameshift\ products} \times 100 \quad (1)$$

The identity of the three bases after stop codons (referred to as +4 to +6) does affect the efficiency of termination, and the decoding release factor is in close physical proximity with the stop codon and positions +4 to +6 of the mRNA [2]. This suggests that the identity of bases in positions +4 to +6 may directly affect release factor recognition. Three series of experiments were carried out to test:

1. what effect do different sequences at positions +4 to +6 have on termination efficiency at RF1 decoded UAG stop codons?

2. what effect do different sequences at positions +4 to +6 have on termination efficiency at RF2 decoded UGA stop codons?

3. do nucleotides at positions +7 to +9 have any effect on termination efficiency?

The sequences investigated were all 64 possible combination of UAG NNN (where N is any nucleotide), 47 of the 64 possible UGA NNN combinations and 42 UGA CUU NNN constructs. The termination efficiency of each construct was determined at least four times.

4 Experimental Data

Experiments were designed for each of the experimentally derived data sets described above.

The data for each termination codon was divided into a training and testing data set, in a ratio of 3:1. The number of examples in each set were as in Table 1. Statistical parameters of the efficiencies in each set are in Table 2. The division of the data sets was done so that the statistical parameters of the training and test set for each group were as similiar as possible.

Codon	Training	Test
UAG	48	16
UGA	36	11
UGACUU	32	10

Table 1: Training and Testing examples available per codon

Data Set	Max	Min	Mean	σ
UAG train	74	34.1	50.4	9.6
UAG test	69.8	35.6	51.2	9.8
UGA train	66.1	10.8	43.5	1.4
UGA test	61.6	18.6	44	1.25
UGACUU train	52.9	15.5	32	7.3
UGACUU test	38.8	22	32.1	5.4

Table 2: Data set statistical parameters

Two different encoding schemes were initially tested. The first represented each base according to their chemical structure, as either a purine (A and G) or a pyrimidine (U and C). Two bits were thus required for this scheme. The second represents each base separately, and thus used four bits. Two different encoding schemes were investigated in an attempt to answer the following question: is the termination efficiency affected by the family of the bases present (i.e. purine versus pyrimidine) or by the individual bases themselves?

5 Initial Experiments

Three neuron layer multi-layer perceptron (MLP) networks were trained for each of the termination codons and for both two-bit and four-bit representations. Performance over the training and testing data sets was evaluated using the R^2 measure (Equation 2).

$$R^2 = \frac{\sum_i^n p_i^2 - \sum_i^n (p_i - t_i)^2}{\sum_i^n p_i^2} \quad (2)$$

where:
p_i is the predicted value of element i
t_i is the target value of element i
n is the number of elements in the target set

After some experimentation, the optimal number of hidden neurons used in the networks, for the four bit representation, were two for UAG and UGA codons, and one for UGACUU. For the two bit representation, one hidden neuron was found to be optimal. Standard backpropagation with momentum

training was used, with the parameters as in Table 3.

Epochs	1000
Learning rate	0.5
Momentum	0.5

Table 3: Training parameters

5.1 Results

The R^2 values across the testing data sets are displayed in Tables 4 and 5 for two and four bit representations, respectively. Plots of the network testing performance are shown in Figures 1, 2 and 3, where the target values are represented by "o", the output values of the two bit networks by "x" and the output values of the four bit network by "+".

Inspection of the tables and plots shows that for the UAG and UGA codons, the four bit representation gave superior results. While the R^2 value for UGACUU codon was high for the two bit representation, the plot indicates that it is a poor performer. Four bit representations were therefore used in the later experiments.

These results allowed us to infer that the four bit representation is superior to the two bit representation. This indicated that the identity of the nucleotides was important in determining termination efficiency rather than the base structures. The results also raise a question: are the inferior results for the UGACUU complex due to the fewer examples available for that codon, or due to the different biological processes involved? This question is addressed in Section 7.

Codon	R^2
UAG	0.973
UGA	0.909
UGACUU	0.964

Table 4: Test R^2 values per codon for 2-bit representation

Codon	R^2
UAG	0.995
UGA	0.981
UGACUU	0.956

Table 5: Test R^2 values per codon for 4-bit representation

6 Sensitivity Analysis

The goal of this research is to gain a better understanding of the problem. Of interest, therefore,

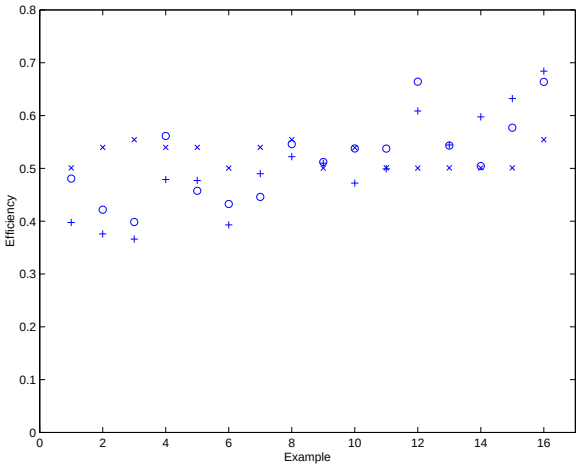

Figure 1: Initial accuracy of UAG networks

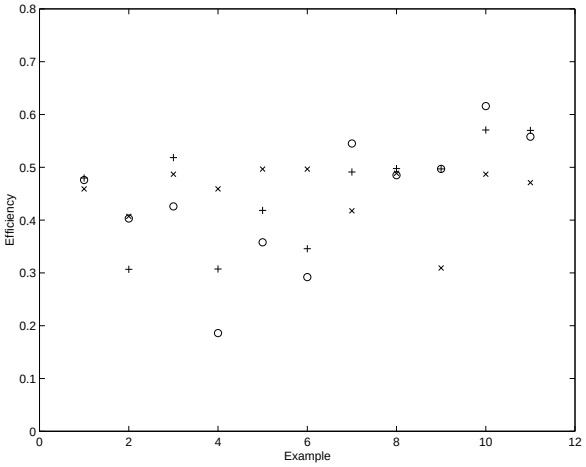

Figure 2: Initial accuracy of UGA networks

is the contribution each input feature (base) makes to the accuracy of the networks. This can be determined by measuring the sensitivity of each input, that is, how much the error of the network is affected by each input feature.

The sensitivity S_j of the network to input feature j is given by Equation 3.

$$S_j = \frac{\sum_i^n E(x_i) - E(\overline{x}_{i,j})}{n} \quad (3)$$

where:
$E(x_i)$ is the network error over example x_i
$E(\overline{x}_{i,j})$ is the network error over example x_i where feature j is set to the mean value of j across all examples
n is the number of examples

The results of this formula are positive if the input is not significant to the network, and negative if it is. The sensitivity of each feature for each network is listed in Table 6.

Each of the features for the UAG and UGACUU networks had negative sensitivities, and thus were

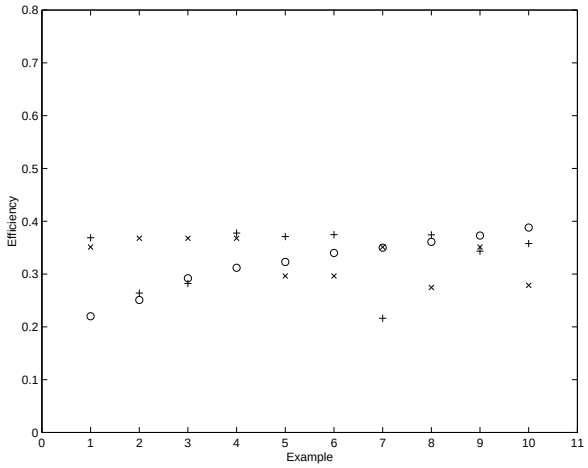

Figure 3: Initial accuracy of UGACUU networks

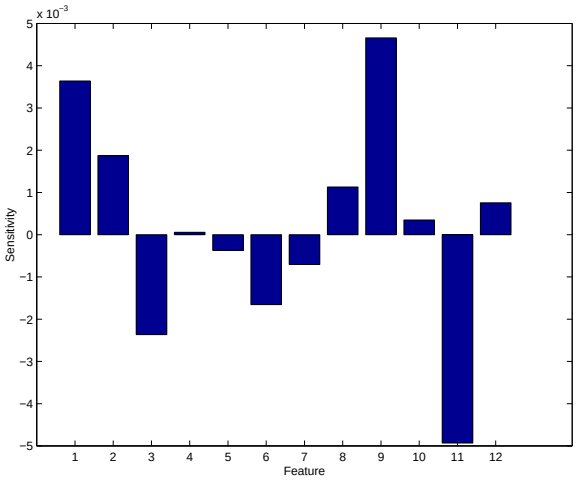

Figure 4: UGA Network Feature Sensitivity

Input	Base	UAG	UGA	UGACUU
1	U	-0.0137	0.0036	-0.0166
2	C	-0.0103	0.00187	-0.0166
3	G	-0.016	-0.00236	-0.0162
4	A	-0.0112	5.65×10^{-5}	-0.0165
5	U	-0.017	-0.000372	-0.0166
6	C	-0.0148	-0.00166	-0.0164
7	G	-0.0135	-0.000706	-0.0164
8	A	-0.0133	0.00113	-0.0165
9	U	-0.013	0.00466	-0.0166
10	C	-0.0145	0.000349	-0.0165
11	G	-0.015	-0.0049	-0.0164
12	A	-0.0122	0.000766	-0.0165

Table 6: Sensitivities

considered important to the network. Five of the features for the UGA network had positive features, which indicated that they are of less importance than the others. Plotting the sensitivity measures on the column graph in Figure 4 showed that the two highest measurements are for the first and third uracil base in the codon.

If these two features are indeed unimportant to the network, then removing them either individually or together should not affect the accuracy of the resulting network. Firstly, the first uracil (U1) was removed from the training and test data sets and a network trained and evaluated. Another network was evaluated with the third uracil (U3) removed, while a third had both U1 and U3 removed from the training and test data sets.

The R^2 values for these three networks across their respective test data sets are displayed in Table 7.

Plots of the performance of each network, where "o" is the target output value and "x" is the actual output value, are displayed in Figures 5, 6 and 7.

Inspection of the R^2 values and plots showed

Removed Base	R^2
U1	0.982
U3	0.986
U1 & U3	0.97

Table 7: R^2 values for UGA networks with removed input features

that the network performance was not decreased by the removal of either the first or third uracil from the data sets, and was only slightly decreased by removal of both. This strongly indicates that the presence or absence of either of these bases is not a major determinant for the ANN model of the UGA stop codons efficiency.

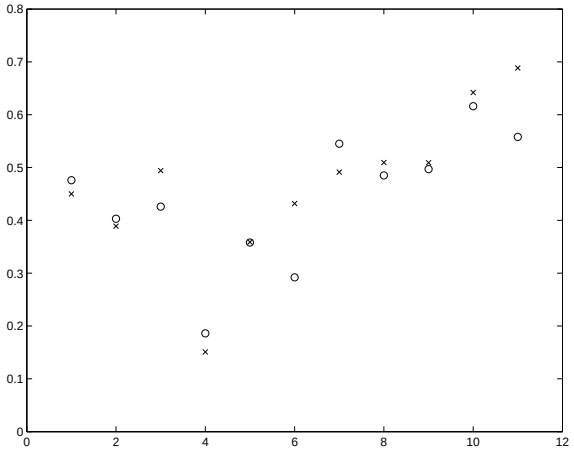

Figure 5: UGA network with base U1 removed

This finding, however, directly contradicts the biological evidence, which strongly suggests that the presence of uracil in the first position consistently leads to a higher termination efficiency.

Modifying the formula used to measure the sensitivity of each feature (Equation 3) to that shown

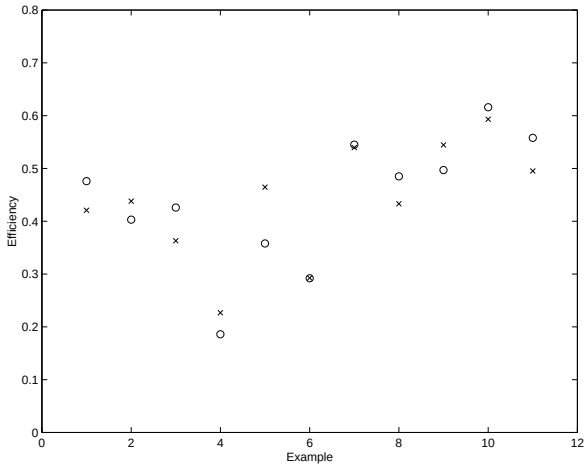

Figure 6: UGA network with base U3 removed

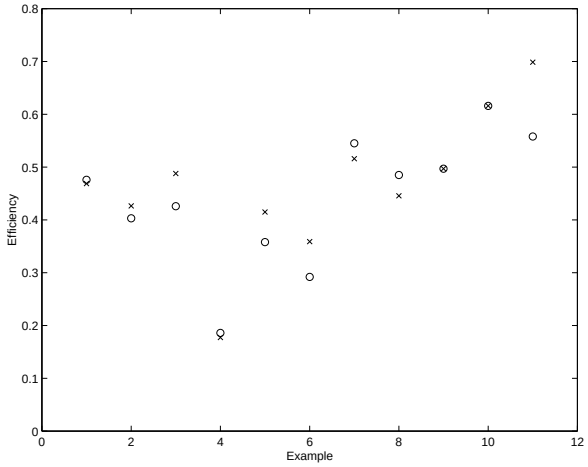

Figure 7: UGA network with both U1 and U3 bases removed

in Equation 4, and reevaluating the sensitivity of each feature of the UGA network, yielded the results shown in Figure 8. It can be seen that the sensitivity measure for the first uracil is higher than that for the other features. At first glance, this would seem to contradict both the sensitivity measures previously determined and the experimental evidence that shows the removal of this feature does not adversely affect the accuracy of the network. A more likely explanation, however, is that the presence or absence of uracil in the first position does not act as a discriminator for termination efficiency.

This interpretation does not contradict the biological data, nor does it contradict the experimental evidence presented here.

$$S_j = \frac{\sum_i^n \mid E(x_i) - E(\overline{x}_{i,j}) \mid}{n} \qquad (4)$$

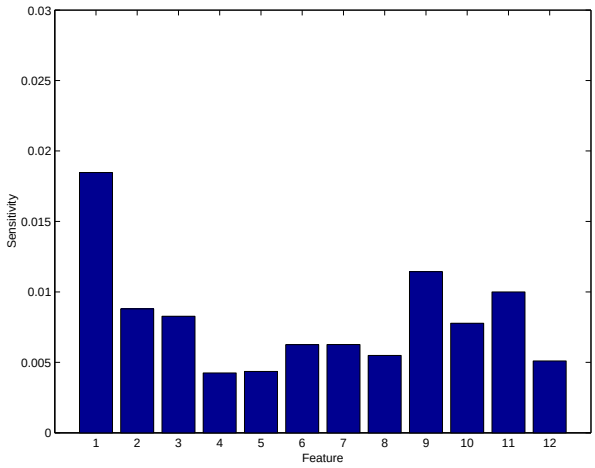

Figure 8: UGA Network Feature Sensitivity

7 Experiments with Reduced Data Sets

As discussed in Section 5, the performance of the networks trained for the UGACUU complex was inferior to that of both UAG and UGA codons.

One of the possible reasons for this performance gap is the smaller number of examples in the UGACUU data sets. To test if this was the case, the training and recall data sets for the UAG and UGA codons had examples removed so that each was the same size as those for UGACUU. Training these networks as before yielded the results in Table 8 and Figures 9 and 10. In these plots, the target values are indicated by "o" and the actual output values by "x".

If the lesser performance of the UGACUU network had been due to the lesser number of examples, then it would be expected that the networks for UAG and UGA would also suffer a drop in performance when the size of their data sets was reduced. Comparison of the R^2 values and plots for the reduced data set UGA and UAG networks with the four bit UGACUU network from Section 5 shows that this is not the case: even with the smaller data sets, the UAG and UGA codon networks perform much better than the UGACUU network. Unless the examples removed from the UAG and UGA data sets were coincidentally (and improbably) examples that are not of great significance to the problem, this result indicates that there truly are different biological processes at work for UGACUU.

Codon	R^2
UAG	0.996
UGA	0.986

Table 8: Test R^2 values per codon for reduced data sets

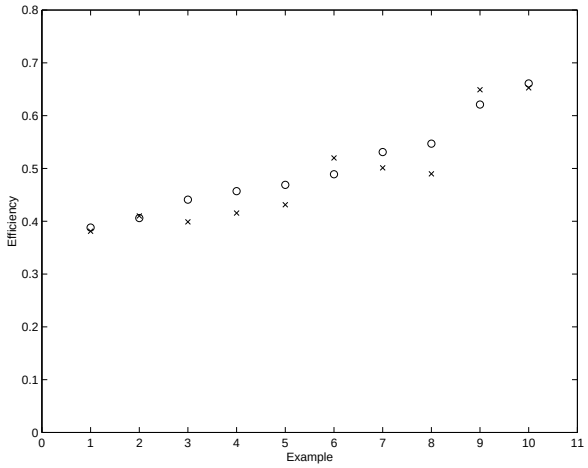

Figure 9: accuracy of UAG network for reduced data sets

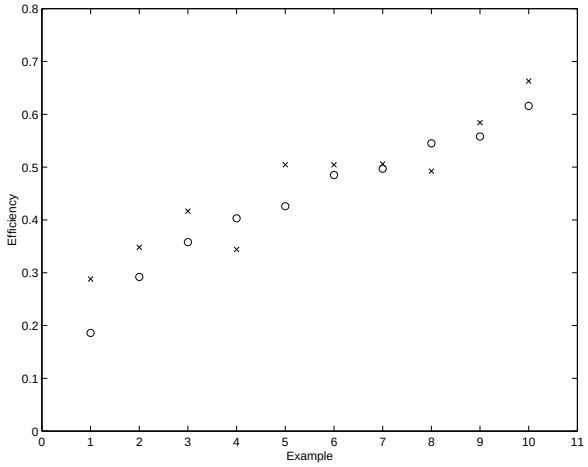

Figure 10: accuracy of UGA network for reduced data sets

8 Future Work

There are several further avenues of research that may be pursued in future. Firstly, a more complete examination of the affects of each input feature is possible. The small size of the networks, and the relatively small number of input features suggests that an exhaustive investigation is possible for each of the codons. As more data becomes available, it will also be possible to carry out further experiments. The question tentatively answered in Section 7 can be more definitively answered when the complete data set is available for the UGACUU codon. Application of one of the several methods of rule extraction from ANN available may also be informative, as it could yield rules that more clearly articulate the functionality of the extended termination signals. Of particular interest is a comparison of the performance of UAG and UGA networks over data for the UAA codon. Since both the RF1 and RF2 proteins decode the UAA codon, it will be interesting to see, when data for UAA becomes available, how accurately the UAG and UGA networks can generalise to it. If the same biological functions underlie the action of RF1 and RF2 for both codons they are keyed to, then the UAG and UGA networks could be expected to generalise to the UAA data as well.

9 Conclusion

It has been shown in Section 5 that MLP are able to successfully model the underlying functions in this problem domain. The superior performance of the four bit representation scheme as opposed to the two bit scheme also evaluated, indicates that the identity of individual bases is of greater importance than their family. The sensitivity analysis performed in Section 6 indicates that for the codon UGA, the presence or absence of uracil in the first position of the trailing codon is not significant as a discriminator between low and high termination efficiencies. The presence of uracil in the third position has been shown to be not significant. Finally, the work performed in Section 7 leads to a tentative conclusion that different biological mechanisms are at work for the second triplet downstream from the stop.

10 Acknowledgements

The authors would like to acknowledge the assistance of Tina Edgar and Po Yee Yip for technical assistance. Some of this work was supported by grants from the Marsden Fund of New Zealand and the Health Research Council of New Zealand to W.P.T.

References

[1] L. L. Major, E. S. Poole, M. E. Dalphin, S. A. Mannering, and W. P. Tate. Is the in-frame termination signal of the *Escherichia coli* release factor-2 frameshift site weakened by a particularly poor context? *Nucleic Acids Research*, 24:2673–2678, 1996.

[2] E. S. Poole, L. L. Major, S. A. Mannering, and W. P. Tate. Translational termination in *Escherichia coli*: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Research*, 26:954–960, 1998.

[3] E.S. Poole, C.M. Brown, and W.P. Tate. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli. EMBO*, 14:151–158, 1995.