Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 28 (2004) 245–256
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Predicting the efficiency of UAG translational stop signal through studies of physicochemical properties of its composite mono- and dinucleotides

Pei-Lin Mao [b], Tie-Fei Liu [b,c], Kelly Kueh [a], Ping Wu [a,*]

[a] *Institute of High Performance Computing, 1 Science Park Road, #01-01 The Capricorn, Singapore 117528, Singapore*
[b] *Institute of Bioengineering and Nanotechnology, 51 Science Park Road, #01-01/10, The Aries, Singapore 117586, Singapore*
[c] *School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore*

Received 4 March 2004; received in revised form 27 May 2004; accepted 29 May 2004

Abstract

In this study, we explored the problem of predicting the UAG stop-codon read-through efficiency. The reported nucleotide sequences were first converted into physicochemical property vectors before being presented to a machine learning algorithm. Two sets of physicochemical properties were applied: one for mononucleosides (in terms of steric bulk, hydrophobicity and electronics) and another for dinucleotides. To the best of our knowledge, this is the first report of how dinucleotides are converted into principle components derived from NMR chemical shift data. A few efficiency prediction models were then derived and a comparison between mononucleoside and dinucleotide-based models was shown. In the derived models, the coefficients of these property based predictors lend themselves to bio-physical interpretations, an advantage which is demonstrated in this study via a prediction model based on the steric bulk factor. Although it is quite simple, the steric bulk factor model explained well the effect of sequence variations surrounding the amber stop codon and the tRNA bearing UCCU anticodon. We further proposed new alternatives at position −1 and +4 of a UAG stop codon sequence to enhance the readthrough efficiency. This research may contribute to a better understanding of the readthrough mechanisms and may also help to study the normal translation termination process.

Keywords: Artificial intelligence; Computational biology; Computer modelling; Physicochemical properties; Amber stop codon

1. Introduction

Translation termination, which is a crucial step in maintaining the life of organisms, is recognized by one of three stop codons (UAG [amber], UAA [orchr], or UGA [opal]), whose efficiency is largely dependent on the context of the codon (Björnsson et al., 1996; Bossi, 1983; Mottagui-Tabar et al., 1994; Poole et al., 1998; Stormo et al., 1986). In readthrough, however, a stop codon is misread as a sense codon and this results in the synthesis of an extended polypeptide. Studies on termination contexts in different cells (see review, Bertram et al., 2001) indicate that the nucleotides immediately after and before the stop codon (defined as −1 and +4) are non-random. Intrinsic physicochemical properties of nucleotides are expected to play a key role in this "programmed translational error", since only normal interactions between the mRNA and components of the translational machinery are involved and no specific gene products have been implicated (Cassan and Rousset, 2001). Little progress, however, has been made in determining the readthrough efficiency from first principles. A ribosome alone contains millions of atoms, which is out of the scope of a molecular dynamic simulation even on the fastest supercomputers. Most of the reported theory-related readthrough models are therefore based on statistical analysis of biological bases (A, U, G, and C) and they do not link to the molecular physics of nucleotides. In the study by Stormo et al. (1986), multiple regressions using Miller and Albertini's (1983) 42 sequence data sets together with an additional 43rd sequence from Bossi (1983) were done (Table 1), which led to a few models showing which bases were important for efficiency. Stormo et al. (1986) as well as others (Major et al., 1996; Poole et al., 1998) found that the nucleotides at +4 (A or G), +5 and −1 (A) play a key

* Corresponding author. Tel.: +65 6419 1212; fax: +65 6778 0522.
E-mail address: wuping@ihpc.a-star.edu.sg (P. Wu).

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2004.05.003

Table 1
Readthrough efficiency data from Miller and Albertini (1983)

Amber site	β-Galactosidase activities[a] (% wild type) *su2*	Amber context −3 −2 −1 +1 +2 +3 +4 +5 +6 N N N u a g N N N
O12	2.3	AUU u a g UCU
A27	3.0	ACC ″ UCC
A9	2.5	AAA ″ UCG
A10	3.8	CAG ″ UUG
O17	0.8	GCG ″ CGC
X15	0.7	GAU ″ CGC
O25	1.8	GCG ″ CGG
O14	2.8	CUA ″ CGA
O18	1.7	AAC ″ CCG
A23	1.2	AUU ″ CCG
A19	3.8	CAC ″ CAA
O27	2.5	UUU ″ CAA
O9	5.1	GCA ″ CAA
O13	11.3	GAU ″ CUG
O10	12.3	CAA ″ CUG
O36	10.3	GUC ″ CUG
A11	8.1	UCG ″ CUG
O35	7.5	CUG ″ CUC
O23	11.5	AAA ″ CUC
O15	8.3	GUC ″ GCC
A6	12.9	AAC ″ GCC
O19	9.3	GUG ″ GCU
A34	8.3	GGC ″ GCG
A20	12.3	GUC ″ GCG
A33	10.6	UCU ″ GGC
X3	18.4	CAA ″ GGU
A18	14.2	GCA ″ GGU
O30	10.6	GGA ″ GAC
A29	10.4	GGA ″ GAC
A15	13.8	GAC ″ GAU
A31	12.1	AAA ″ GAU
O11	18.0	UCG ″ AUU
O24	20.7	AAU ″ AUU
O21	14.6	CAG ″ AUC
A30	25.6	UGU ″ AUC
A5	9.4	UAU ″ ACC
O28	10.1	CAA ″ ACC
O34	9.2	GGG ″ ACC
A16	9.6	GAC ″ ACA
A24	10.8	GAC ″ AGU
A13	19.3	GUG ″ AUG
A26	17.8	GAU ″ AUG
A17	1.3[b]	GUG ″ CAU

[a] The numbers are the percent activity of β-galactosidase when the amber codon is suppressed by *su2* as compared to the non-amber (wild type) codon.
[b] The activity of A17 is from Bossi (1983).

role in the readthrough efficiency, but no physicochemical properties were put forward to explain the phenomenon of suppression due to limitations of the mathematical approach using biological bases as the model elements.

Recently, a number of materials research projects were successfully carried out using correlation techniques (Heng et al., 1999; Jin et al., 2000; Wu and Heng, 1999; Wu et al., 1999, 2002) to predict a bulk material property from the fundamental atomic properties of the constituent elements. These models may lead to approximated physics of otherwise very complicated compound formation mechanisms since all atomic properties are well known (or can be easily computed by well established quantum mechanics). Without correlation approaches, it is very difficult if not impossible in many cases to derive the mechanisms from first principles methodology alone.

Similar efforts were reported to link sequence activity to physicochemical properties of its composite nucleotides (Jonsson et al., 1993; Sjöström et al., 1986), in which principal component analysis was performed on a data set of 21 experimentally determined and calculated nucleoside properties (Sandberg and Sjöström, 1996). Four (4) statistically significant components, or principal properties (*P*) were extracted which described 68.4% of the variance in the data. Since the principal properties are condensed descriptors from the original property data set, each of the four *P*s can be related to physicochemical properties of the nucleotides; specifically these are: P_1 relates to the steric bulk, P_2 relates to the hydrophobicity, P_3 relates to the electronic properties, and P_4 relates to the electronic/hydrophobic properties. A typical steric bulk property (P_1) is the heat of formation: −91, −127, −176 and −230 kcal/mol for nucleoside A, G, C and U, respectively. Other useful steric bulk properties include molecular weight and total molecular surface area. Hydrophobicity (P_2) is well represented by the logarithm of octanol/water partition coefficient: −2.83, −2.88, −3.08 and −3.85 for U, A, C, and G, respectively. Electronic properties (P_3) may be obtained from the energy of lowest unoccupied molecular orbital: −0.538, −0.29, −0.064 and −0.032 eV for G, A, U and C, respectively. G has the strongest electron affinity while C has the weakest. Based on these four *P*s and the developed sequence correlation model, Sandberg and Sjöström (1996) proposed and experimentally verified a new sequence with high activity of gene expression. They also explained the physical bases on the nucleotides that are favoured at certain sequence positions.

In the present study, the four *P*s proposed by Sandberg and Sjöström are used to develop mononucleoside based readthrough models. For dinucleotide based readthrough models, we constructed principal properties (pps) for dinucleotides using the same procedure as for the mononucleosides. A challenge in constructing pps is finding sufficient dinucleotide data from the literature since a property can only be used if data is available for all 16 different dinucleotides. After an extensive literature search, we found seven parameters of proton chemical shifts (NMR), 1‴, 2′, 2″, 3′, 4′ 5′, 5″ (Cheng and Sarma, 1977) for all 16 different dinucleotides under five different experimental conditions. Therefore, we had a total of 560 (16 × 7 × 5) data points, which were used in this study to derive the principal properties for dinucleotides. Although this data uses deoxyribodinucletotides instead of ribodinucleotides that are directly involved in readthrough, we cannot find any experimental information on the 16 ribodinucleoside monophosphates. The difference between thymine and uracil is, however, limited to the C5 position (–CH_3 group in thymine and –H

in uracil), in which only one proton NMR is affected. The only other site affected is in the sugar group. Thus, the 16 deoxyribodinucleoside monophosphates also may still reveal the intrinsic intermolecular interactions of the 16 ribodinucleoside monophosphates.

In this work, the three pps were used to construct dinucleotide based readthrough models. We compare the two groups of models and try to link the observed readthrough efficiency to physical aspects of the nucleotides. Lastly, a new steric bulk model is proposed to explain some observed effects of sequence variations surrounding the amber stop codon and the tRNA with UCCU anticodon.

2. Methods

2.1. APEX (advanced process expert)

An in-house developed pattern recognition software tool, APEX (Jin et al., 1999; Wu et al., 2004, 2002), is used to derive correlation models. Interested readers may repeat the model development by applying MATLAB (Hunt et al., 2001) and the flow charts of APEX (Jin et al., 1999; Wu et al., 2002), which involves data preprocessing, collinearity checking, feature reduction and pattern recognition. Principal component analysis (PCA), partial least squares regression (PLSR), cross-validation and PRESS (prediction residual error sum of squares) calculations are the important techniques used. In particular, PCA is used for extracting the principal properties and PLSR is used to establish quantitative relationship models. A brief introduction to PCA and PLSR is given below.

2.2. Principal component analysis (PCA)

PCA transforms the original parameters into a new set of uncorrelated variables called principal components. It aims to obtain the principal component score vectors (principal properties), which are used to understand the physicochemical differences among nucleotides, and the loading vectors, which describe the physicochemical nature of the score vectors.

$$X_{np} = S_{nf} \times F_{fp} + E_{np} \quad (1)$$

where X is the original spectral data, S the PCA scores, F the PCA factors (eigenvectors, loadings), E the matrix of residual spectral, n the number of samples, p the number of data-points and f the number of principal components.

PCA breaks apart the spectral data into the most common spectral variations (factors, eigenvectors, and loadings) and the corresponding scaling coefficients (scores).

2.3. Partial least squares regression (PLSR)

Partial least squares regression is a linear regression method that forms components (factors, or latent variables) as new independent explanatory variables in a regression model. As with every regression method, the regression model from PLSR can be expected to have a smaller number of components without an appreciably smaller R-square value.

The linear PLSR model is

$$Y = \sum_{i=1}^{n} \lambda_i X_i + \varepsilon \quad (2)$$

where Y is a vector of suppression activity for each sequence, X a matrix of principal properties for each nucleotide, λ the coefficient vector calculated and ε the constant (error) vector of the model.

Since most of the biological data is generally skewed, pre-processing of the data is needed in order to get a good fit. A logarithm is used to normalize Y and the X matrix is also normalized before computing the λ values of the model.

3. Results

This article is organised in four parts. First, correlation models based on mononucleosides will be presented, followed by construction of the three principal components from dinucleotide properties. Next, we will present the correlation models based on dinucleotides and lastly we will discuss the results.

3.1. Mononucleotide-based models

Four principal properties, P_1, P_2, P_3 and P_4 are taken directly from Sandberg and Sjöström (1996) to describe the four nucleosides as shown in Table 2. According to Sandberg, these four significant principal properties represent steric bulk (P_1), hydrophobic (P_2), electronic (P_3), and both electronic and hydrophobic (P_4) descriptor variables. Therefore, mononucleotide A is mainly characterized by its large size and high electronegativity, C by high electronegativity, G by high hydrophilicity, and U by a balance in size, hydrophobic and electronic property (Table 2).

A standard multivariate regression was performed using the readthrough efficiency data in Table 1. For all 43 sequences (each has nine nucleotides of which three nucleotides compose the amber codon), each nucleotide is replaced by its four Ps except the amber codon, which are not included in the model. Take the amber site O12 for an example; the sequence [A U U u a g U C U] is replaced

Table 2
Principal properties of mononucleosides (N)

N	P_1	P_2	P_3	P_4
A	2.25	0.77	−1.17	0.75
C	−2.36	0.72	−2.07	−1.76
G	−0.31	−3.76	0.03	1.41
U	−3.96	1.61	−1.72	0.53

by [2.25 0.77 −1.17 0.75, −3.96 1.61 −1.72 0.53, −3.96 1.61 −1.72 0.53, u a g, −3.96 1.61 −1.72 0.53, −2.36 0.72 −2.07 −1.76, −3.96 1.61 −1.72 0.53]. Therefore, a prediction model for mononucleotides (full model) is obtained [$\log(\text{activity}) = \sum_{i,j} a_{i,j} P_{i,j} + \varepsilon$]:

$$\begin{aligned}\log(\text{activity}) &= -0.0373P_{1,-3} + 0.0028P_{2,-3} - 0.0225P_{3,-3} \\ &\quad -0.0186P_{4,-3} - 0.0153P_{1,-2} - 0.0016P_{2,-2} \\ &\quad -0.0522P_{3,-2} - 0.0526P_{4,-2} + 0.0961P_{1,-1} \\ &\quad +0.0945P_{2,-1} - 0.1310P_{3,-1} - 0.0054P_{4,-1} \\ &\quad +0.2049P_{1,+4} - 0.0772P_{2,+4} + 0.2106P_{3,+4} \\ &\quad +0.0578P_{4,+4} - 0.1603P_{1,+5} + 0.1712P_{2,+5} \\ &\quad -0.1132P_{3,+5} + 0.2899P_{4,+5} - 0.0123P_{1,+6} \\ &\quad +0.0282P_{2,+6} + 0.0042P_{3,+6} + 0.079P_{4,+6} + 1.7520 \end{aligned} \quad (3)$$

where $P_{i,j}$ is the property i at position j and $a_{i,j}$ is its coefficient. The amber codon nucleotides, at position +1, +2 and +3, are not in the model. Fig. 1 shows the observed versus model predicted suppression activities. This full model has a R^2 of 89.56%, which indicates that the correlation between mononucleoside properties and the suppression activities is strong. The validity of the model as a predictive tool is assessed through the "leave-one-out" cross-validation using the function *plscross* in MATLAB (Hunt et al., 2001).

Similar to Sandberg's approach, the regression coefficients of the mononucleoside based model are plotted for each principal property against the sequence positions as shown in Fig. 2. It is clear that position +5, +4 and −1 are very important since they have the largest coefficients for almost every property. A submodel with just these three positions has also been derived. The coefficients of the properties are listed in Table 3. It has a R^2 value of 87.18%, indicating that most of the information from the full model is contained in these three positions.

The term ($|a_{i,-1}|+|a_{i,+4}|+|a_{i,+5}|$) may measure the contribution of each P_i to the submodel and a rank of P_i is obtained: P_1 (0.45) > P_3 (0.42) >P_4 (0.39) >P_2 (0.33). The steric bulk (P_1) is therefore the most important factor to readthrough efficiency.

3.2. Dinucleotide-based models

A natural extension from the mononucleoside model is to include intrinsic properties related to the 16 different pairs of nucleotides, since physicochemical properties related to the linkage of two nucleotides are ignored in the mononucleoside models. Principal component analysis was conducted on the 560 proton chemical shift data (NMR) of dinucleotides. We constructed three (3) statistically significant principal components, pp1, pp2 and pp3 that accounted for 97.2% of the variation in the 560 data points, whereas, pp1 carried the most variation with 80.2%, pp2 the second with 12.0%, and pp3 with 5.0%. The loading plots shown in Fig. 3 indicated that pp1 was dominated by proton chemical shifts of 1′, 3′, 4′, 5′ and 5″, pp2 by 2′, whereas, pp3 was dominated by 2″, 3′ and 4′. According to Xu et al. (1998), ^{13}C chemical shifts of nucleotides are strongly correlated to their backbone torsion angles. It is reasonable to expect that the above proton chemical shifts are also related to some structure parameters of dinucleotides though the exact relationship is not yet known. Therefore, the three pps have their roots in specific structural features of the dinucleotides. Table 4 shows the dinucleotide principal properties for each nucleotide pair; Fig. 4 is the resultant score plot of the principal component analysis.

3.3. Dinucleotide-based models using principal properties

Models based on the principal properties shown in Table 4 were developed with PLSR using the readthrough efficiency data (Table 1). As in mononucleoside based models, all 43 data sets (each has nine nucleotides of which three nucleotides compose the amber codon) were transferred into the sequence of dinucleotide properties; each nucleotide pair

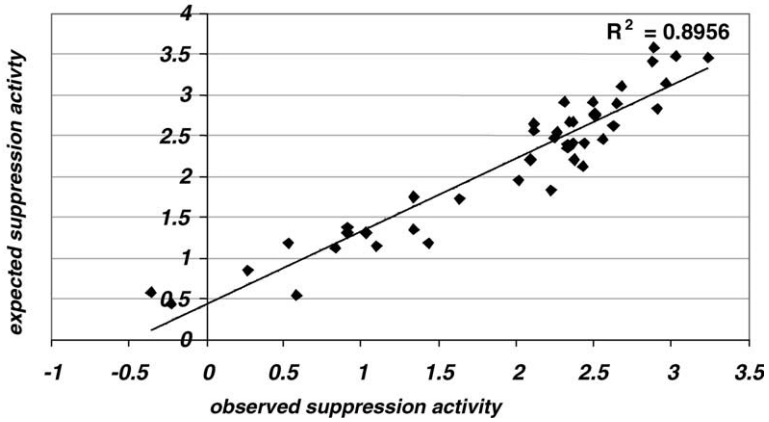

Fig. 1. PLS correlation plot for the two-dimensional model, showing the amber suppression activity from the mononucleoside model (expected suppression activity) vs. the corresponding literature data (observed suppression activity). Suppression activity is logarithmic.

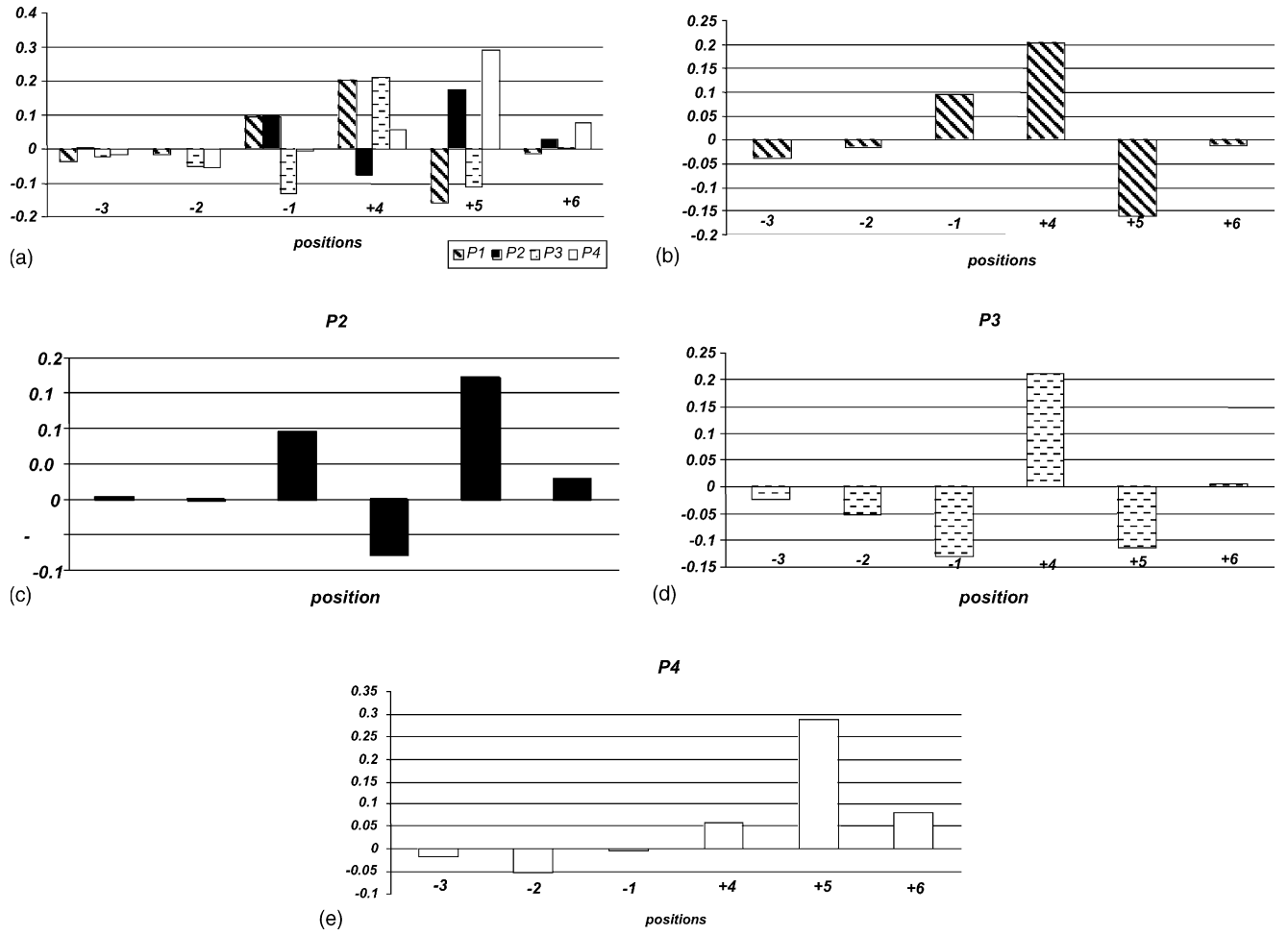

Fig. 2. (a) Principle properties of mononucleoside effect and their influence on the activity of six positions before and after UAG displayed as PLS regression coefficients; (b) coefficients for steric bulk (P_1) vs. sequence position; (c) coefficients for hydrophobicity (P_2) vs. sequence position; (d) coefficients for electron affinity (P_3) vs. sequence position and (e) coefficients for electronegativity/hydrophobicity (P_4) vs. sequence position.

is replaced by its three pps except two pairs within the amber codon, which are not included in the model. Using the amber site O12 as an example, the sequence (A, U, U, u, a, g, U, C, U) is shown as (AU, UU, Uu, ua, ag, gU, UC, CU) and can be replaced by (0.370 2.158 0.868, 0.954 −0.448 0.430, 0.954 −0.448 0.430, u a, a g, 0.812 0.651 0.864, 0.996 −0.319 0.406, −0.050 −0.775 −0.611).

When developing an optimal dinucleotide based prediction model, the optimal number of principal components needed in the model and the validity of the model as a predictive tool are assessed using "leave-one-out" cross-validation; this can also be done by MATLAB (Hunt et al., 2001) using the function *plscross*. APEX (Jin et al., 1999; Wu et al., 2004, 2002) calculates the optimal number of principal component as three. The number of factors needed for regression is then determined and used to compute the PLS regression coefficients. An optimal model is thus developed [$\log(\text{activity}) = \sum_{i(\text{m,n})} a_{i,(\text{m,n})} \text{pp}_{\text{i(m,n)}} + \varepsilon$], where $\text{pp}_{\text{i,(m,n)}}$ is the property i for the pair located at (m, n) and $a_{i,(\text{m,n})}$ is the coefficient. In Fig. 5, the observed and predicted suppression activities using this equation (Poole et al., 1998) are plotted against each other with an R^2 of 84.22%, indicating that the principal properties of dinucleotides have a large impact on the suppression activity. The activity of each sequence can be calculated using other models in Table 3.

Examining the 18 (3 × 6) regression coefficients by each property, as shown in Fig. 6, clearly shows the highest peak at position (+3, +4) for all properties. The second highest peak is at position (+4, +5) for properties 1 and 3, followed by a distant third at position (−1, +1). Similarly, several models using different combinations of positions were obtained in a similar procedure. Clearly, positions (+3, +4) and (+4, +5) account for most of the information (near 82%) and (−1, +1) is another possibly significant position. The prediction model using these three dinucleotides (with $R^2 = 84.03\%$) is shown in Table 3.

4. Discussion

In the literature, the observed readthrough efficiency is determined using an empirical rule for tetranucleotides as the

Table 3
Coefficients of PLSR models for mono- and dinucleotides

Full PSLR model				Sub PSLR models			
Mononucleotide[a]		Dinucleotide[b]		Mononucleotide[a]		Dinucleotide[b]	
Property	Coefficient	Property	Coefficient	Property	Coefficient	Property	Coefficient
$P_{1,-3}$	−0.0373	$pp_{1,(-3,-2)}$	−0.1046	$P_{1,-1}$	0.0827	$pp_{1,(-1,+1)}$	0.09999
$P_{2,-3}$	0.0028	$pp_{2,(-3,-2)}$	0.1299	$P_{2,-1}$	0.0686	$pp_{2,(-1,+1)}$	0.1408
$P_{3,-3}$	−0.0225	$pp_{3,(-3,-2)}$	−0.1711	$P_{3,-1}$	−0.1056	$pp_{3,(-1,+1)}$	0.2263
$P_{4,-3}$	−0.0186	$pp_{1,(-2,-1)}$	0.0020	$P_{4,-1}$	−0.0278	$pp_{1,(+3,+4)}$	17.3399
$P_{1,-2}$	−0.0153	$pp_{2,(-2,-1)}$	0.0830	$P_{1,+4}$	0.1849	$pp_{2,(+3,+4)}$	1.2587
$P_{2,-2}$	−0.0016	$pp_{3,(-2,-1)}$	−0.2473	$P_{2,+4}$	−0.0956	$pp_{3,(+3,+4)}$	1.2158
$P_{3,-2}$	−0.0522	$pp_{1,(-1,+1)}$	−0.2154	$P_{3,+4}$	0.2222	$pp_{1,(+4,+5)}$	1.2935
$P_{4,-2}$	−0.0526	$pp_{2,(-1,+1)}$	0.2038	$P_{4,+4}$	0.0436	$pp_{2,(+4,+5)}$	0.0182
$P_{1,-1}$	0.0961	$pp_{3,(-1,+1)}$	−0.2252	$P_{1,+5}$	−0.1797	$pp_{3,(+4,+5)}$	0.3655
$P_{2,-1}$	0.0945	$pp_{1,(+3,+4)}$	14.4488	$P_{2,+5}$	0.1664	Constant[c]	−9.5415
$P_{3,-1}$	−0.1310	$pp_{2,(+3,+4)}$	−1.1574	$P_{3,+5}$	−0.0879	–	–
$P_{4,-1}$	−0.0054	$pp_{3,(+3,+4)}$	−1.1446	$P_{4,+5}$	0.3222	–	–
$P_{1,+4}$	0.2049	$pp_{1,(+4,+5)}$	−1.54	Constant[c]	1.8564	–	–
$P_{2,+4}$	−0.0772	$pp_{2,(+4,+5)}$	0.0471	–	–	–	–
$P_{3,+4}$	0.2106	$pp_{3,(+4,+5)}$	−0.2765	–	–	–	–
$P_{4,+4}$	0.0578	$pp_{1,(+5,+6)}$	0.0094	–	–	–	–
$P_{1,+5}$	−0.1603	$pp_{2,(+5,+6)}$	−0.042	–	–	–	–
$P_{2,+5}$	0.1712	$pp_{3,(+5,+6)}$	0.083	–	–	–	–
$P_{3,+5}$	−0.1132	Constant[c]	−7.07413	–	–	–	–
$P_{4,+5}$	0.2899	–	–	–	–	–	–
$P_{1,+6}$	−0.0123	–	–	–	–	–	–
$P_{2,+6}$	0.0282	–	–	–	–	–	–
$P_{3,+6}$	0.0042	–	–	–	–	–	–
$P_{4,+6}$	0.079	–	–	–	–	–	–
Constant[c]	1.7520	–	–	–	–	–	–

[a] For the models of mononucleotides, $\log(\text{activity}) = \sum_{i,j} a_{i,j} P_{i,j} + \varepsilon$ $P_{i,j}$ is the property i at position j and $a_{i,j}$ is its coefficient.

[b] For the models of dinucleotides, $\log(\text{activity}) = \sum_{i(\text{m,n})} a_{i(\text{m,n})} \text{pp}_{\text{i(m,n)}} + \varepsilon$ where $\text{pp}_{\text{i(m,n)}}$ is the property i for the pair located at (m, n) and $a_{i(\text{m,n})}$ is the coefficient.

[c] ε is the constant.

fourth base controls the efficiency of termination (Bonetti et al., 1995; Poole et al., 1995), or by some near cognate tRNA (Yarus and Curran, 1992). The nature of the flanking 3′ base was first shown to be important in experiments with *lacI–lacZ* fusions in *Escherichia coli*, whose data (Miller and Albertini, 1983) has been applied in this paper. The suppressor tRNA performed more efficiently when the amber codon was flanked with A or G (UAGA or UAGG) than with U or C (UAGC or UAGU). Studies on the effect of the nucleotides on the 5′ end of the termination codon indicated that the last two N-terminal amino acids have influence in *E. coli* (Björnsson et al., 1996; Mottagui-Tabar et al., 1994; Mottagui-Tabar et al., 1997). Both RF-1 and RF-2 (Craigen et al., 1990) associate with a third factor, RF-3 (Grentzmann et al., 1994; Mikuni et al., 1994), for maximal efficiency to release the nascent peptide from the P-site tRNA (Crawford et al., 1999; Freistroffer et al., 1997; Grentzmann et al., 1995). Thus, the P-site tRNA and its interaction with its codon or the release factors (Zhang et al., 1996) and possibly also some modified bases in the P-site tRNAs (Yarus and Curran, 1992) are important factors in the efficiency of termination.

Based on these experimental observations, it is commonly accepted that nucleotides at −1, +4, and possibly +5 and +6 are critical to readthrough efficiency. For –1, there is a gradation in the readthrough efficiency by which A > G, and at +4, it is A > U. We shall discuss the performance of the mononucleotide and the dinucleotide based models.

Table 4
PCA score value of each dinucleotide

	pp1	pp2	pp3
AA	1.030	−0.585	0.640
AG	0.746	−0.192	0.733
AC	0.819	0.848	0.911
AU	0.370	2.158	0.868
GA	0.748	−0.813	0.396
GG	0.762	−0.398	0.569
GC	0.770	0.609	0.761
GU	0.812	0.651	0.864
CA	0.905	−1.653	0.211
CG	0.922	−1.394	0.310
CC	0.950	−0.412	0.463
CU	−0.050	−0.775	−0.611
UA	0.983	−1.706	0.105
UG	0.945	−1.343	0.243
UC	0.996	−0.319	0.406
UU	0.954	−0.448	0.430

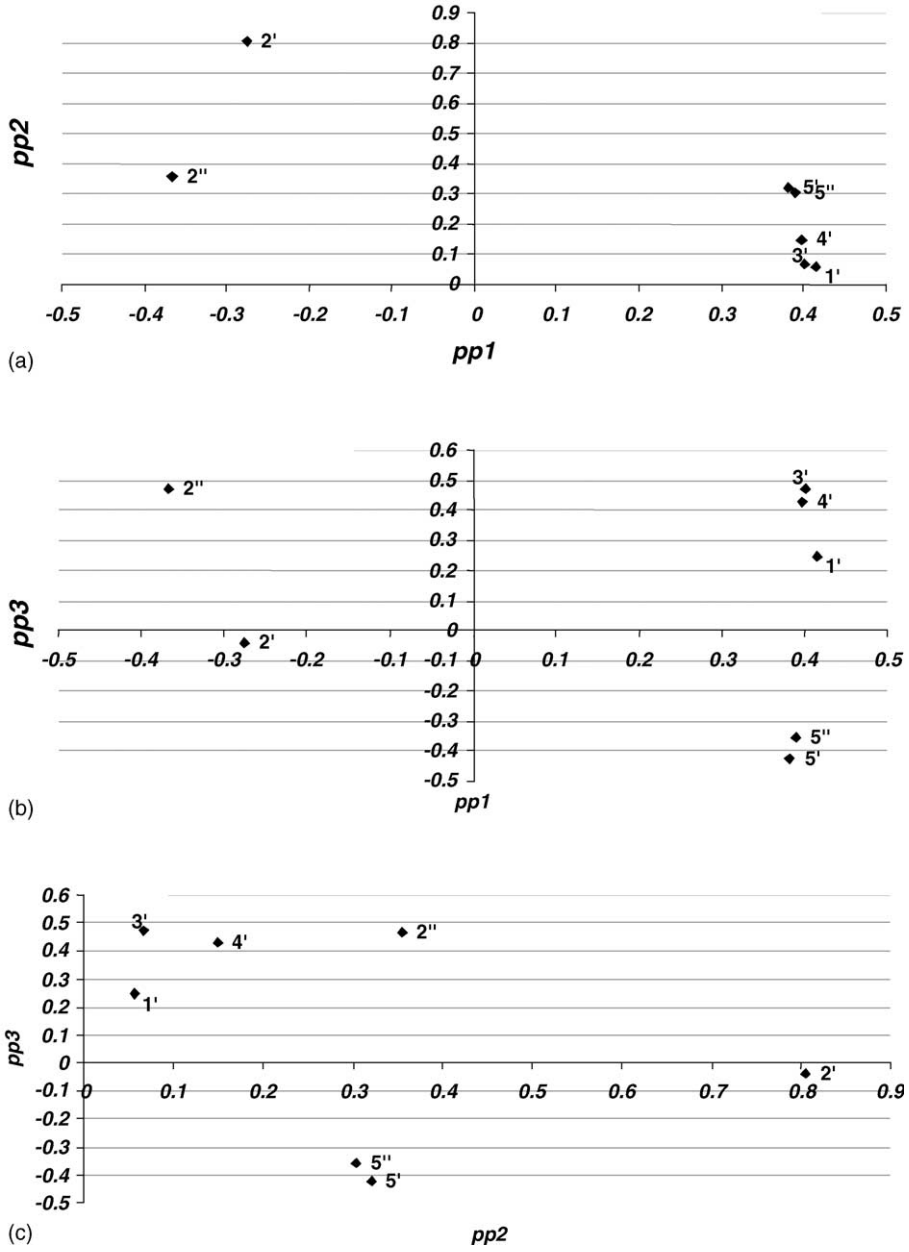

Fig. 3. PCA loading plots of dinucleotide's principle properties (pp). (a) pp1 vs. pp2; (b) pp1 vs. pp3 and (c) pp2 vs. pp3.

4.1. Mononucleotide-based model

As shown in Fig. 2, +5, +4 and −1 have the highest regression coefficients for most of the four *P*s, so that those positions contribute significantly to the readthrough efficiency. This prediction is consistent with experimental observations. The sign of a *P* matters when comparing its value at different positions, which may indicate the trend of its impact at across all positions, or among the four *P*s at a fixed position, which may show the interdependence of all four *P*s at a certain position.

Contributions of each *P* in accordance with all six positions (−3, −2, −1, +4, +5, +6) are shown in Fig. 2. For −1, contribution of the sum of four *P*s is 0.44, 0.12, −0.01, and −0.40, respectively for A, C, U and G. Bases A and C may be preferred for this position with A > C. By similar analysis at +4, G and A may be the most important contributors to high readthrough efficiency. The model predicted site preference of the −1 and +4 sites agrees well with experiments. The site preference of −3, −2, +5, and +6, is also predicted, respectively, as U, C, U and U. Therefore, two sequences of UCAUAGGUU and UCAUAGAUU may have high efficiency readthrough. Eight (8) recently reported sequences (Zhang et al., 1999) were used to test the model reliability. The selected eight sequences, GGNUAGUGU and AUNUAGUGU with N = A, C, G or U, were designed

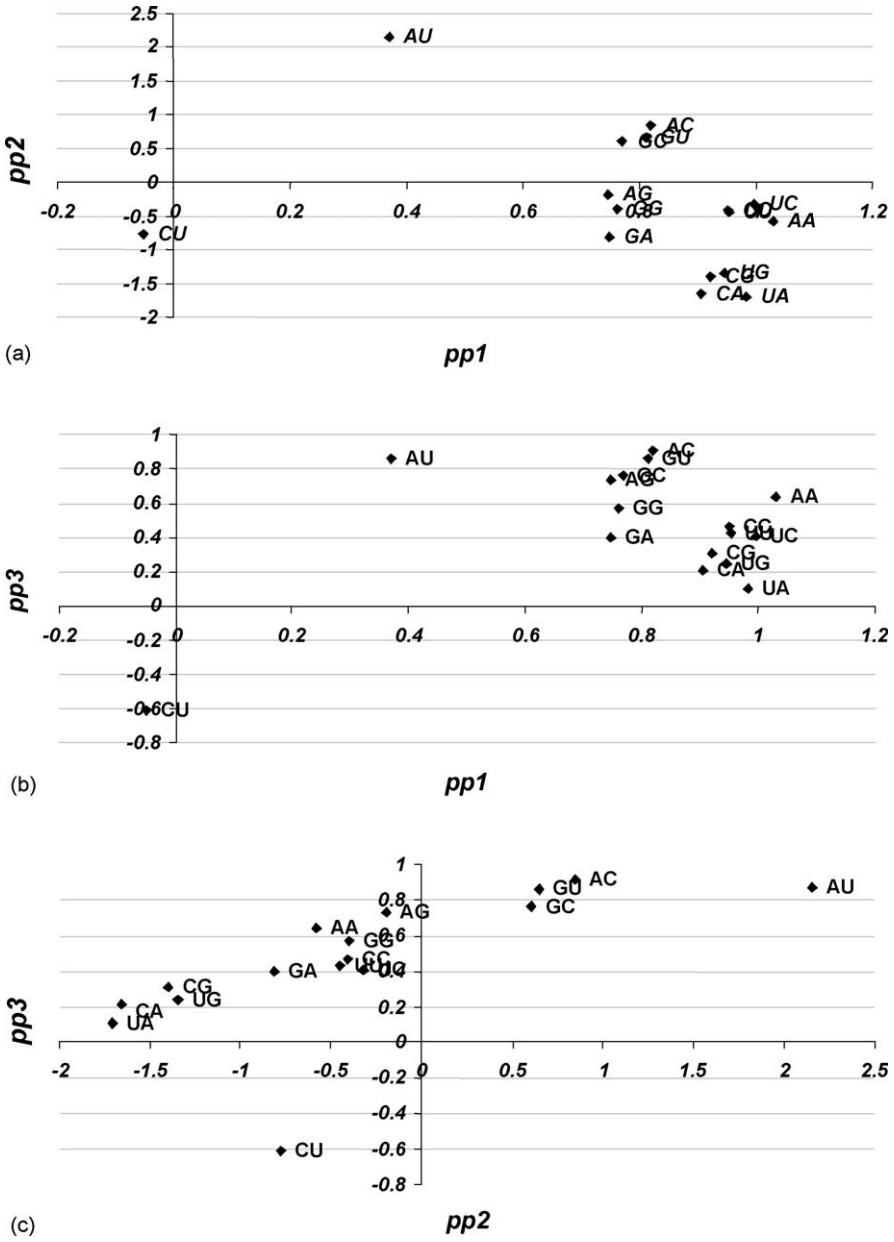

Fig. 4. Score plot of dinucleotide principal properties for PCA model. (a) pp1 vs. pp2; (b) pp1 vs. pp3 and (c) pp2 vs. pp3.

to test the −1 site preference. The reported transmission for GGNUAGUGU decreases as N = A (0.61), G (0.54), C (0.25), and U (0.23), while the model predicted trend is A, C, U and G. The model only failed on the N = G sequence. Similarly, for AUNUAGUGU, the readthrough efficiency decreases in the order: A (0.20), C (0.18), G (0.16) and U (0.10). The model also missed the N = G sequence. The dinucleotide models will fix this problem by considering bonding characteristics between two nucleotides.

From Fig. 2, the −1 site involves size (P_1), hydrophobic (P_2) and electronic (P_3) factors. A preferred nucleotide should have large size (P_1), small hydrophobicity (P_2) and large electron affinity (P_3). It calls for the modified nucleotide, 1-methyladenosine, m^1A ($P_1 = 3.87$, $P_2 = -0.32$ and $P_3 = -1.64$). Similarly, the modified nucleotide, 6-inosineadenosine, i^6A ($P_1 = 6.32$, $P_2 = 3.29$ and $P_3 = 0.96$) is preferred at +4 site. Another possible preferred base for +4 is the modified nucleotide, 7-methylguanosine, m^7G ($P_1 = 1.42$, $P_2 = -4.41$, and $P_3 = 0.76$) of strong hydrophilic. New experiments may be of interest to test the sequence of NNm^1AUAGi^6ANN or NNm^1AUAGm^7GNN to verify these predictions.

4.2. Dinucleotide-based models

As shown in Fig. 6(a), the highest regression coefficients were for the (+3, +4), (+4, +5), and (−1, +1) pairs [with a predominant pp1 for the (+3, +4) pair], so that the +4, +5

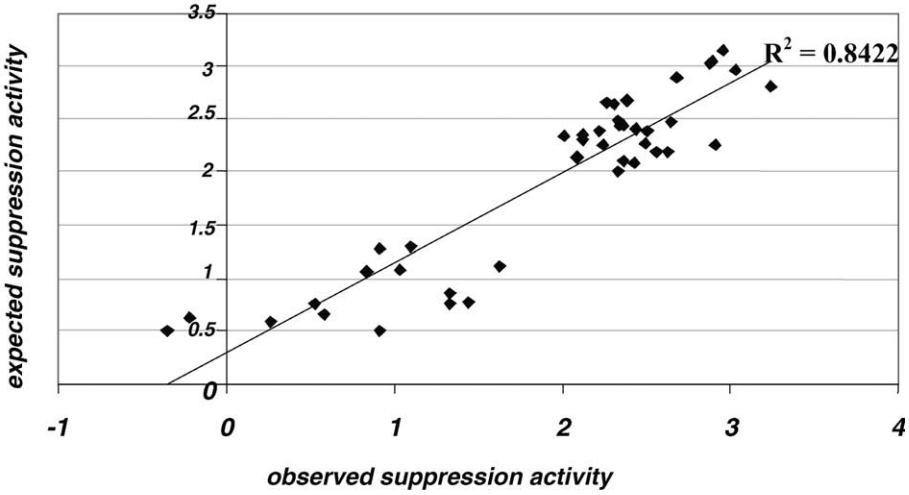

Fig. 5. PLS correlation plot for the two-dimensional model, showing the amber suppression activity from the dinucleotide model (expected suppression activity) vs. the corresponding literature data (observed suppression activity). Suppression activity is logarithmic.

and −1 positions contribute significantly to the readthrough efficiency. The predictions of the model agree with experiments. In looking at +4, G is always at +3 since this is the last position of the amber codon and thus the (+3, +4) pair has to start with G. Using a similar argument as in the mononucleotide based model, bases A and G are preferred with A > G > C > U. At the −1 site, the predicted preferred base is A and G with A > G, which is in agreement with experiments.

Since the (+3, +4) pair is mainly characterized by pp1 (which was dominated by proton chemical shifts of 1′, 3′, 4′, 5′ and 5′), the structural features that are described by proton chemical shifts of 1′, 3′, 4′, 5′ and 5″ may be responsible or linked to the molecular interaction mechanism involving the (+3, +4) pair during the readthrough process. Similar to the mononucleotide model, eight (8) recently reported sequences (Zhang et al., 1999), GGNUAGUGU and AUNUAGUGU with N = A, C, G or U, are used to test the dinucleotides model reliability. The predicted order is N = A > G > C > U for the GGN series and N = A > G > C > U for the AUN series. The model predictions agreed well with experiments for GGN and AUN, except the order of AUG and AUC where the experimental data are very close (C (0.18) and G (0.16)). The dinucleotides models allow a close examination of bonding relationships among different bases, and may lead to better predictions than the mononucleotide models. The sequence AUAGA is, therefore, expected for high readthrough efficiency and will be used to study possible physical features of the amber stop codon readthrough.

By using pattern recognition techniques and experimental readthrough efficiency data, two groups of readthrough prediction models were developed using physicochemical properties of mono- as well as dinucleotide properties. Comparisons of model predictions with experiments are summarized in Table 5. Since the mononucleotide based models demonstrated that the size factor P_1 of nucleotides is the dominant control factor governing readthrough efficiency, it is expected that the size factor may account for most of the observed contextual gradients that affect the efficiency of readthrough. The size factor (P_1) is a condensed descriptor from the original 21 properties and is strongly related to the size of molecules. To simplify the discussion, one of the original size properties is selected to represent P_1. The heat formation is one of the typical steric bulk properties and will be applied to explore possible relationships between the sequence of steric bulk and readthrough efficiency. As indicated earlier, among the four bases, A has the largest size and the highest heat of formation (−91 kcal/mol), G has the second highest (−127 kcal/mol), C the third (−176 kcal/mol), and U the lowest (−230 kcal/mol). Among the four possible URR sequences with R = A or G, there are three stop codons excluding UGG. Since U has the lowest formation energy and A has the highest, a stop codon may be characterized by the largest energy gap (from low to high, 230 −

Table 5
Comparison of models and experiments on site significance and its occupancy

Model index	Experimental observed significant sites	Model predicted significant sites	Experimental site occupancy	Model predicted site occupancy
Mononucleotides	+4, +5, −1, +6	+5, +4, −1	+4 (A > U)	+4 (G > A > U)
	–	–	−1 (A > G)	−1 (A > G)
Dinucleotides	+4, +5, −1, +6	+4, +5, −1	+4 (A > U)	+4 (A > G > C > U)
	–	–	−1 (A > G)	−1 (A > G > C > U)

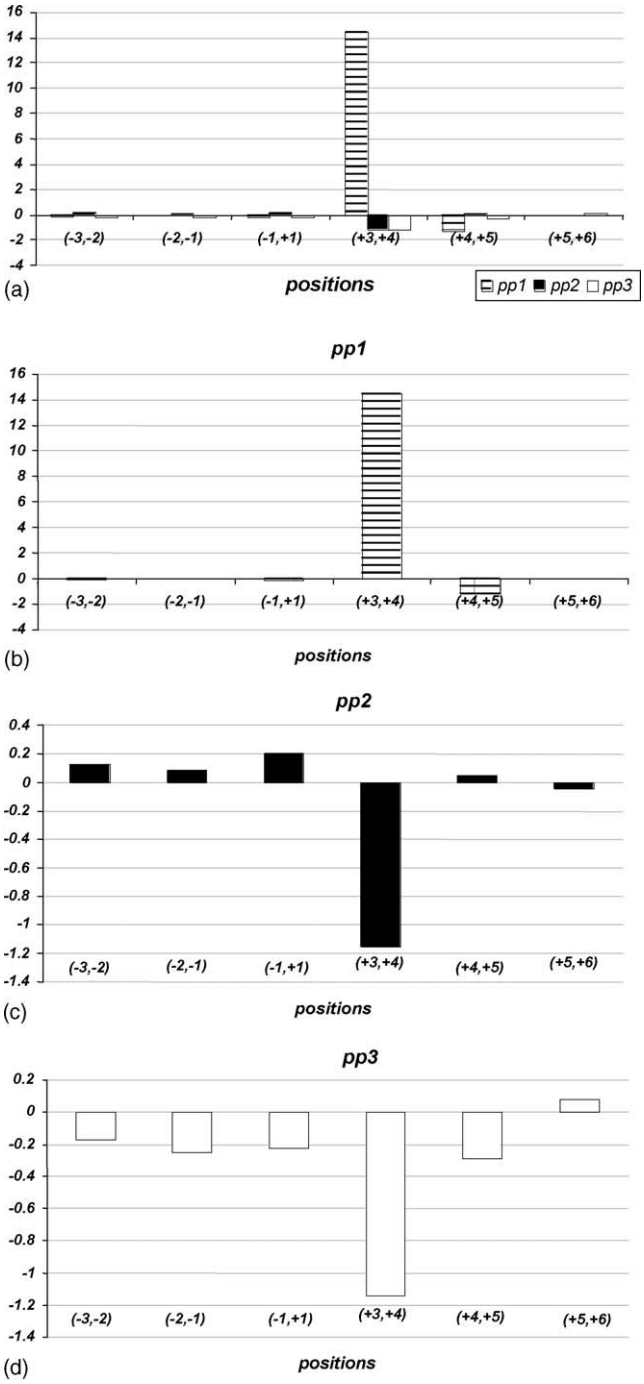

Fig. 6. Principle properties of dinucleotide effect and their influence on the activity of six positions before and after UAG codon in considered seven different nucleotide pairs (shown in figure) displayed as PLS regression coefficients (a). (b) pp1 in each nucleotide pair; (c) pp2 in each nucleotide pair and (d) pp3 in each nucleotide pair.

91 = 139 kcal/mol) between +1 and either +2 or +3. Although G has the second largest energy, UGG is not a stop codon. It is further found that UAA is decoded by both RF-1 and RF-2 and is the most frequently used stop codon in *E. coli*. (Brown et al., 1990; Martin et al., 1988), which may indicate the importance of the sharp energy contrast (or a large size difference) shown in UAA compared to UAG or UGA.

It is interesting to use the steric bulk properties to examine the AUAGA sequence which is predicted to have high readthrough efficiency. The sequence has the lowest energy (U) at the +1 site then rises up to the highest (A) at +2 and lowers a little to the second highest (G) at +3. As indicated above, A is preferred when connecting to +1 (U) from −1. Since A is a high energy state, this leads to a large energy contrast at the start of a stop codon, which may help to amplify the energy contrast within the stop codon [between +1 (U) and +2 (A)]. Similarly, when moving on from +3, a high-energy state (A) is also preferred at +4. It seems a small energy difference between +3 and +4 is helpful in ending the process, which may also serve to smooth the energy plateau of the stop codon. This observation agrees with the fact that A and G (A > G) are preferred at +4 for all three types of stop codons (URR, R = A or G, excluding UGG). This is because these codons have either A or G at +3 and the energy plateau is best extended if +4 is also occupied by either A or G. Experimentally, UAAU is shown to be the strongest and most efficient four base stop signal (or the least readthrough) in the CAI subset with the highest expression (Tate et al., 1996) (where codon adaptation index (CAI), is a measure of sense codon usage correlating with expression). This agrees with the steric bulk argument since the UAAU has the largest energy gap (or size difference) surrounding the end of the stop codon (U at +4 is the worst smoother). Similarly, UAAA should be the most efficient readthrough codon, which has the smallest energy gap surrounding the end of the stop codon (A at +4 as the best smoother).

Readthrough efficiency is the result of competition between termination mediated by release factors (RFs) and elongation mediated by a cognate termination tRNA or a suppressor tRNA (readthrough) in a complex biomolecule system with ribosomes and elongation factor Tu (EF-Tu). Therefore, the above proposed steric bulk model may be over simplified for such a complex system; nevertheless, the model is capable of explaining many readthrough observations. In the four-base codon AGGN family, for example, the series of readthrough efficiency appears to be AGGA > AGGG > AGGC ∼ AGGU with a variety of efficient tRNA suppressors containing UCCU anticodons from a library (Magliery et al., 2001). This agrees with the steric bulk factor model since the energy of A or G at +4 is closer to that of G at +3 (good smoothers), while C or U at +4 is further away (bad smoothers). The steric bulk model may also be linked to some observations in codon–anticodon interaction research in which efficient suppressors of four-base codon AGGA were obtained. To study the effect of sequence variation surrounding the UCCU anticodon from the $tRNA^{Ser}$(N8) library, different amounts of ampicillin (20 μg ml^{-1} to 1500 μg ml^{-1}) were applied (Magliery et al., 2001). They reported that at 1500, 300 and 20 μg ml^{-1} ampicillin, the sequences converged, respectively, on 5′-MUUCCUAM-3′, 5′-NBUCCURN-3′, and

5′-NNNCCUDN-3′, where M = A or C, B = C or G or U, R = A or G, D = A or G or U, and N = A or G or C or U. The positions from 5′ to 3′ of the above anticodon sequences are defined sequentially from t1 to t8, respectively. With decreasing ampicillin level from 1500 to 20 $\mu g\,ml^{-1}$, position t7 relaxes from A to G, and further to U, corresponding to a reduction in the energy of the base from −91 (A) to −230 (U). It seems as the energy of the t7 base is decreased, the anticodons become more structurally flexible and more active in the codon–anticodon reaction, and thus leads to the increase in the number of different suppressors. The t7 base on anticodon is close in space to the −1 site in the codon sequence AGGA. Assuming the −1 site prefers a similar local energy environment as the t7 site, the corresponding occupation of −1 site is from A to G and to U just as it was in the t7 site. The size difference between the −1 and +1 (A) sites will then increase (AA < GA < UA) which will enhance the readthrough efficiency based on the steric bulk model. Similarly, when the level of ampicillin is reduced, the t2 site relaxed from U to C or G, and further to any base. This is also consistent with the steric bulk model shown for the amber stop codon. Since t2 is close to the end of the AGGA codon, a small energy gap is expected between the +5 site (which is assumed to have the same occupation base as the t2 site) and +4 (A) site when +5 (or t2) changes from being a U to something else and narrows the energy gap.

More experiments will be useful to further examine and improve the proposed steric bulk model. It is not clear why a large size difference is preferred around the start of the amber codon while a small difference is preferred around the end of the codon. It seems, however, the steric bulk model reproduces many observed trends in the sequence variation of some codons and anticodons in prokaryotes. For systems containing biomolecules outside the training data set of the present study (Namy et al., 2001), their control factors (rather than the size factor) and codon–anticodon interaction mechanisms may be different and new readthrough models may be generated by studying the correlations of readthrough experimental data with the physicochemical properties of the composite mononucleosides and dinucleotides.

Acknowledgements

The APEX software was developed with a grant from the exploratory funding of the National Science and Technology Board of Singapore. We would like to thank Dr. Michael B. Sullivan for a critical English proof reading and useful suggestions.

References

Bertram, G., Innes, S., Minella, O., Richardson, J.P., Stansfield, I., 2001. Endless possibilities: translation termination and stop codon recognition. Microbiology 147, 255–269.

Björnsson, A., Mottagui-Tabar, S., Isaksson, L.A., 1996. Structure of the C-terminal end of the nascent peptide influences translation termination. EMBO J. 12, 1696–1704.

Bonetti, B., Fu, L.W., Moon, J., Bedwell, D.M., 1995. The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in *Saccharomyces cerevisiae*. J. Mol. Biol. 251, 334–345.

Bossi, L., 1983. Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J. Mol. Biol. 164, 73–87.

Brown, C.M., Stockwell, P.A., Trotman, C.N.A., Tate, W.P., 1990. The signal for the termination of protein synthesis in prokaryotes. Nucleic Acids Res. 18, 2079–2086.

Cassan, M., Rousset, J.-P., 2001. UAG readthrough in mammalian cells: effect of upstream and downstream stop codon contexts reveal different signals. BMC Mol. Biol. 2, 3–10.

Cheng, D.M., Sarma, R.H., 1977. Intimate details of the conformational characteristics of deoxyribodinucleoside monophosphates in aqueous solution. J. Am. Chem. Soc. 99, 7333–7348.

Craigen, W.J., Lee, C.C., Caskey, C.T., 1990. Recent advances in peptide chain termination. Mol. Microbiol. 4, 861–865.

Crawford, D.-J.G., Ito, K., Nakamura, Y., Tate, W.P., 1999. Indirect regulation of translational termination efficiency at highly expressed genes and recoding sites by the factor recycling function of *Escherichia coli* release factor RF3. EMBO J. 18, 727–732.

Freistroffer, D.V., Pavlov, M.Y., MacDougall, J., Buckingham, R.H., Ehrenberg, M., 1997. Release factor RF3 in *E. coli* accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. EMBO J. 16, 4126–4133.

Grentzmann, G., Brechemier-Baey, D., Heurgue, V., Mora, L., Buckingham, R.H., 1994. Localization and characterization of the gene encoding release factor RF3 in *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 91, 5848–5852.

Grentzmann, G., Brechemier-Baey, D., Heurgue, V., Mora, L., Buckingham, R.H., 1995. Function of polypeptide-chain release factor RF3 in *Escherichia coli*-RF3 action in termination is predominantly at UGA-containing stop signals. J. Biol. Chem. 270, 10595–10600.

Hunt, B.R., Lipsamn, R.L., Rosenberg, J.M., 2001. A Guide to MATLAB: for Beginners and Experienced Users. Cambridge University Press.

Heng, K.L., Jin, H.M., Li, Y., Wu, P., 1999. Computer aided design of Ni-MH electrodes. J. Mater. Chem. 9, 837–843.

Jin, H.M., Li, Y., Wu, P., 1999. Prediction of new additives for galvanising process. J. Mater Res. 5, 1791–1795.

Jin, H.M., Li, Y., Liu, H.L., Wu, P., 2000. Study on the behaviours of additives in steel hot dip galvanising by DFT calculations. Chem. Mater. 12, 1879–1883.

Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., Wold, S., 1993. Quantitative sequence-activity models QSAM-tools for sequence design. Nucleic Acids Res. 21, 733–739.

Magliery, T.J., Anderson, J.C., Schultz, P.G., 2001. Expanding the genetic code: selection of efficient suppressors of four-base codons and identification of shifty four-base codons with a library approach in *Escherichia coli*. J. Mol. Biol. 307, 755–769.

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

Martin, R., Weiner, M., Gallant, J., 1988. Effects of release factor context at UAA codons in *Escherichia coli*. J. Bacteriol. 170, 4714–4717.

Mikuni, O., Ito, K., Moffat, J., Matsumura, K., McCaughan, K., Nobukuni, T., Tate, W.P., Nakamura, Y., 1994. Identification of the prfC gene, which encodes peptide-chain-release factor 3 of *Escherichia coli* is similar but different from *Salmonella typhimurium*. Proc. Natl. Acad. Sci. U.S.A. 88, 3758–3761.

Miller, J.H., Albertini, A.M., 1983. Effects of surrounding sequence on the suppression of nonsense codons. J. Mol. Biol. 164, 59–71.

Mottagui-Tabar, S., Bjornsson, A., Isaksson, L.A., 1994. The second to last amino acid in the nascent peptide as a codon context determinant. EMBO J. 13, 249–257.

Mottagui-Tabar, S., Bjornsson, A., Isaksson, L.A., 1997. Only the last amino acids in the nascent peptide influence translation termination in *Escherichia coli* genes. FEBS Lett. 414, 165–170.

Namy, O., Hatin, I., Rousset, J.-P., 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep. 2, 787–793.

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

Poole, E.S., Major, L.L., Mannering, S.A., Tate, W.P., 1998. 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 Acid Res. 26, 954–960.

Sandberg, M., Sjöström, M., 1996. A multivariate characterization of tRNA nucleosides. J. Chemometr. 10, 493–508.

Sjöström, M., Wold, S., Söderström, B., 1986. In: Gelsema, E.S., Kanal, L.N. (Eds.), Pattern Recognition in Practice, II. Elsevier, Amsterdam, p. 486.

Stormo, G.D., Schneider, T.D., Gold, L., 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res. 14, 6661–6679.

Tate, W.P., Dalphin, M.E., Pel, H.J., Mannering, S.A., 1996. The stop signal controls the efficiency of release factor-mediated translational termination. Genet. Eng. 18, 157–182.

Wu, P., Heng, K.L., 1999. Correlation of chemical element properties to electrochemical properties of hydrogen storage alloys. Commun. Chem. Mater. 11, 858–861.

Wu, P., Jin, H.M., Li, Y., 1999. Relationship among chemical element properties, bulk additive properties, and crystal structures of binary zinc-compounds. Chem. Mater. 11, 3166–3170.

Wu, P., Jin, H.M., Liu, H.L., 2002. Investigation of additive effects on Zn diffusion in hot-dip galvanizing by DFT calculations. Chem. Mater. 14, 832–837.

Wu, P., Zeng, Y.Z., Wang, C.M., 2004. Prediction of apatite lattice constants from their constituent elemental radii and artificial intelligence methods. Biomaterials 25, 1123–1130.

Xu, X.-P., Chiu, W.-L.A.K., Au-Yung, S.C.F., 1998. Chemical shift and structure relationship in nucleic acids: correlation of backbone torsion angles γ and α with ^{13}C chemical shifts. J. Am. Chem. Soc. 120, 4230–4231.

Yarus, M., Curran, J., 1992. In: Hatfield, D.L., Lee, B.Y., Pirtle, R.M. (Eds.), Transfer RNA in Protein Synthesis. CRC Press, Boca Raton, pp. 319–365.

Zhang, S.P., Rydén-Aulin, M., Isaksson, L.A., 1996. Functional interaction between release factor one and P-site peptidyl-tRNA on the ribosome. J. Mol. Biol. 261, 98–107.

Zhang, S.P., Rydén-Aulin, M., Isaksson, L.A., 1999. Interaction between a mutant release factor one and P-site peptidyl-tRNA is influenced by the identity of the two bases downstream of the stop codon UAG. FEBS Lett. 455, 355–358.