

The use of sequence comparison to detect 'identities' in tRNA genes

Jun-Ichi Sagara*, Seishi Shimizu, Takeshi Kawabata¹, Shugo Nakamura, Mitsunori Ikeguchi and Kentaro Shimizu

Department of Biotechnology, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113, Japan and
¹Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima 411, Japan

Received November 24, 1997; Revised and Accepted March 3, 1998

ABSTRACT

We have developed a computational method that detects 'identities' in tRNA genes by using principal component analysis to classify the sequences of bases in tRNA genes into groups of similar sequences and then comparing the distribution of sequences of bases, in order to extract characteristic bases that are conserved within a group but differ between groups. These classification and comparison procedures are applied recursively to classify the sequences into hierarchical groups, so that multiple levels of characteristic sites can be detected. By using this computational method, we were able to detect many characteristic sites in the T and D domains of tRNAs, as well as the characteristic sites that had already been detected experimentally. This suggests that bases not only in the contact regions but also in the elbow regions, which determine the structure and dynamics of the whole tRNA molecule, are important to the tRNA–aminoacyl tRNA synthetase recognition.

INTRODUCTION

Molecular recognition related to tRNAs is one of the most interesting themes in molecular biology and has been approached from various directions. Aminoacyl tRNA synthetases (ARSs) catalyze the linkage of tRNAs to amino acids (1–4), and are known to recognize sites that consist of a few bases of their cognate tRNAs. These sites, which may include non-anticodon bases, are called 'identities' (5,6), and if we are to understand the mechanism of this recognition, we need to be able to detect identities in various kinds of tRNAs.

A number of experimental approaches for detecting identities in tRNAs have been developed. The recent determination of crystal structures of two complexes between tRNAs and their cognate ARSs, for example, as well as the analysis of mutants of the tRNAs, has made it possible to explain the structure–function relationship of their interactions. Although these approaches have resulted in many identities being detected, we will not have a coherent picture of the structural basis of aminoacylation specificity until we have more experimental data.

To facilitate the identification of determinants of the specificity, we have developed a computational method for identifying 'characteristic sites', nucleotide positions that determine the characteristic interaction of a tRNA. The sequences of tRNA genes in a genome database are first classified into groups by principal component analysis (PCA) of multiple sequence alignment. Gene sequences are represented as vectors in a generalized sequence space, and groups of similar sequences are revealed when these vectors are projected onto a lower-dimensional space. The distribution of bases is then compared with the distribution of sequences by using multidimensional scaling analysis (MDS), in which the bases of each sequence are projected individually onto the same sequence space. This makes it possible to extract characteristic bases for each group.

This method is based on that used by Casari *et al.* (7) to predict functional residues in protein families. The method was used to find out which groups of residues specific to particular subfamilies are responsible for functional differences between protein subfamilies. We extended Casari's method to the analysis of tRNA gene sequences and used it to identify the groups of bases specific to particular tRNA classes. We encoded four types of bases as 4-bit binary numbers and constructed the gene sequence vector. We also applied the above procedures (PCA and MDS) recursively in order to classify the sequence into hierarchical groups. Since the Ras–Rab–Rho superfamily that was analyzed by Casari *et al.* (7) was already known to have three subfamilies, the recursive application did not seem to be necessary. In the analysis of tRNA genes, however, such knowledge was not available beforehand.

There have been several computational approaches to detecting identities in tRNAs by sequence comparison. McClain *et al.* (8), for example, developed a method called the Disjoint Subject (DS) routine for analyzing a single nucleotide position, and Atilgan *et al.* (9) developed a method called the Expectation Maximization (EM) routine for analyzing multiple nucleotide positions. In the EM routine, all possible combinations of any specified number of positions are compared between composite and individual tRNA sequences. These methods are based on the comparison of individual sequences, whereas our method classifies all the input sequences at once into one or more groups by using a simple multivariate analysis method (PCA), and extracts characteristic bases from each group. The advantages of our method are as follows: (i) it allows comparison of the whole sequence and can

*To whom correspondence should be addressed. Tel: +81 3 3813 8728; Fax: +81 3 3813 8723; Email: jun@bi.a.u-tokyo.ac.jp

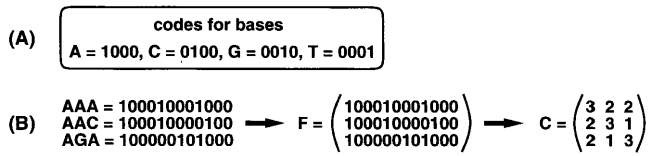


Figure 1. Sequence profiles. (A) Encoding rule. (B) Example of an alignment matrix.

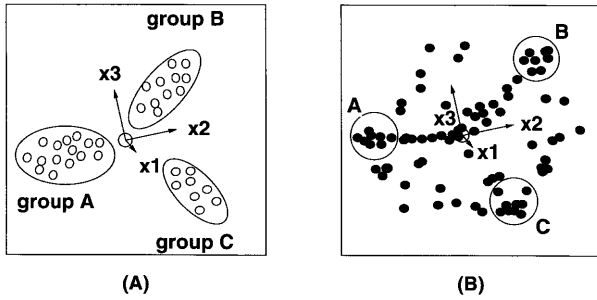


Figure 2. Sequence space. (A) Plots of sequences. (B) Plots of bases.

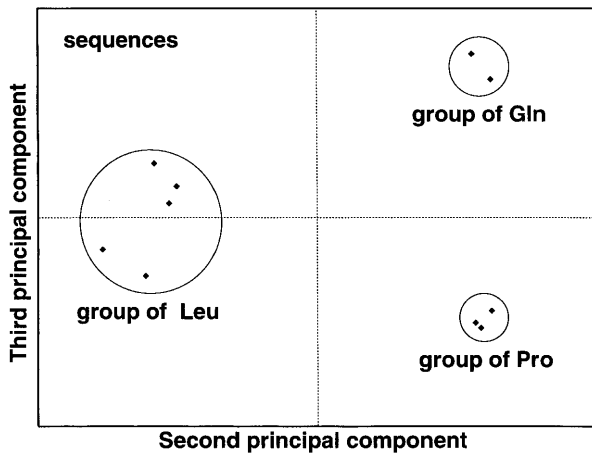


Figure 3. Plots of sequences (tRNA^{Gln}, tRNA^{Pro} and tRNA^{Leu}).

detect many candidate identities; (ii) the recursive application of PCA and MDS makes it possible to classify the sequences into hierarchical groups and to detect multiple levels of characteristic bases; and (iii) the algorithm is simple and its computational cost is small.

Using this method, we found characteristic sites containing many experimentally determined identities. We also found that many of the characteristic sites detected in our method are in the T and D domains. This suggests that not only the bases in the contact regions but also the bases in the elbow region, which determine the structure and dynamics of the whole tRNA molecule, are important to the tRNA-ARS recognition.

MATERIALS AND METHODS

The PCA is first applied to the entire sequences of tRNA genes. We define an alignment matrix F , each row of which is a sequence vector \vec{F}^k for the k th gene sequence. Each base is encoded to a 4-bit binary number (A, C, G and T are respectively encoded to

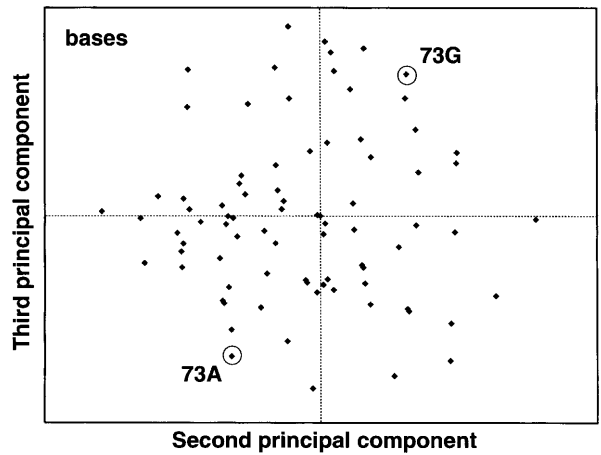


Figure 4. Plots of single sequence bases (tRNA^{Gln}, tRNA^{Pro} and tRNA^{Leu}).

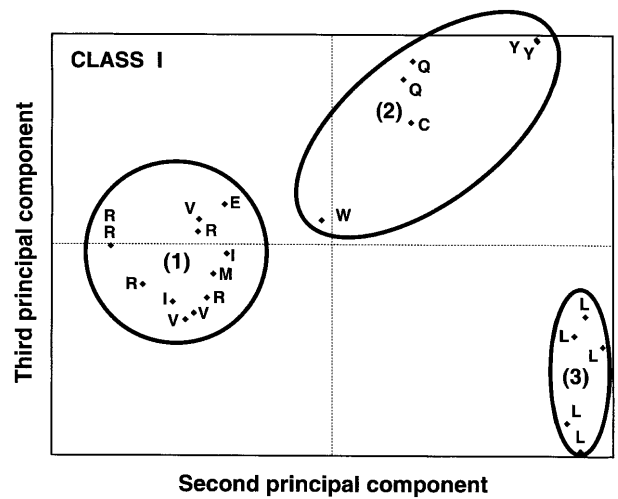


Figure 5. Plots of sequences (tRNAs in CLASS I).

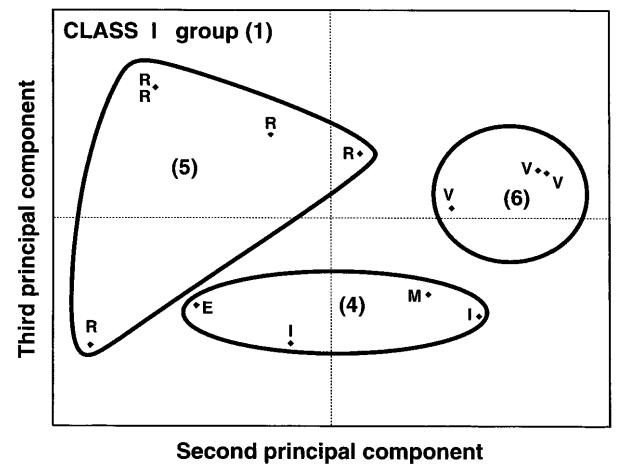


Figure 6. Plots of sequences [recursive analysis for Group (1)].

1000, 0100, 0010 and 0001) (Fig. 1A). A sequence vector consists of 1s and 0s and corresponds to a point in 4l-dimensional space,

where l is the length of the sequence alignment. For n sequences of genes, an alignment matrix is defined as a $4l \times n$ matrix:

$$F = \begin{bmatrix} \vec{F}^1 \\ \vdots \\ \vec{F}^k \\ \vdots \\ \vec{F}^n \end{bmatrix} \quad 1$$

Figure 1B shows an example of an alignment matrix based on the encoding in Figure 1A. The alignment matrix is analogous to conventional profiles derived from multiple alignments. Just as conventional profiles give a tabular summary of the base content at each position in an alignment, each element of the sequence vector is 1 or 0, depending only on whether or not a particular base type exists at a sequence position.

The number $C^{kk'}$ of matched bases between sequences k and k' can be expressed as the inner product of the sequence vectors:

$$C^{kk'} = \vec{F}^k \cdot \vec{F}^{k'} \quad 2$$

A comparison matrix C , each element of which is the number of matches for all pairs of sequences, can thus be expressed as the matrix product between alignment F and its transpose F^T :

$$C = F F^T \quad 3$$

The principal axes \vec{u}_p are defined as

$$C\vec{u}_p = \lambda_p \vec{u}_p \quad 4$$

where \vec{u}_p is an eigenvector and λ_p is the corresponding eigenvalue of comparison matrix C . Each sequence is plotted on the two-dimensional plane called the sequence space. The coordinate x_p^k of gene k in dimension p is given by

$$x_p^k = \sqrt{\lambda_p} u_p^k \quad 5$$

Sequences are classified into one or more groups, according to the distance between the two-dimensional sequence plots.

Then, the MDS is applied. Bases of each sequence are projected individually onto the same two-dimensional plane in order to trace the principal components back to individual bases and positions that characterize individual groups. The coordinates \vec{y}_p of bases in a sequence are given by

$$\vec{y}_p = F^T \vec{u}_p \quad 6$$

The i th element of \vec{y}_p corresponds to a base at position i in the sequence, and characteristic bases of each group are detected by comparing the bases with the groups of sequences. Applying the classification and comparison procedures above recursively, enables the groups to be classified into subgroups, and the characteristic bases in the subgroups to be found. The recursive application makes the results of the classification clearer.

Figure 2 shows a schematic description of the sequence space to illustrate our method. In Figure 2A, sequences (open circles)

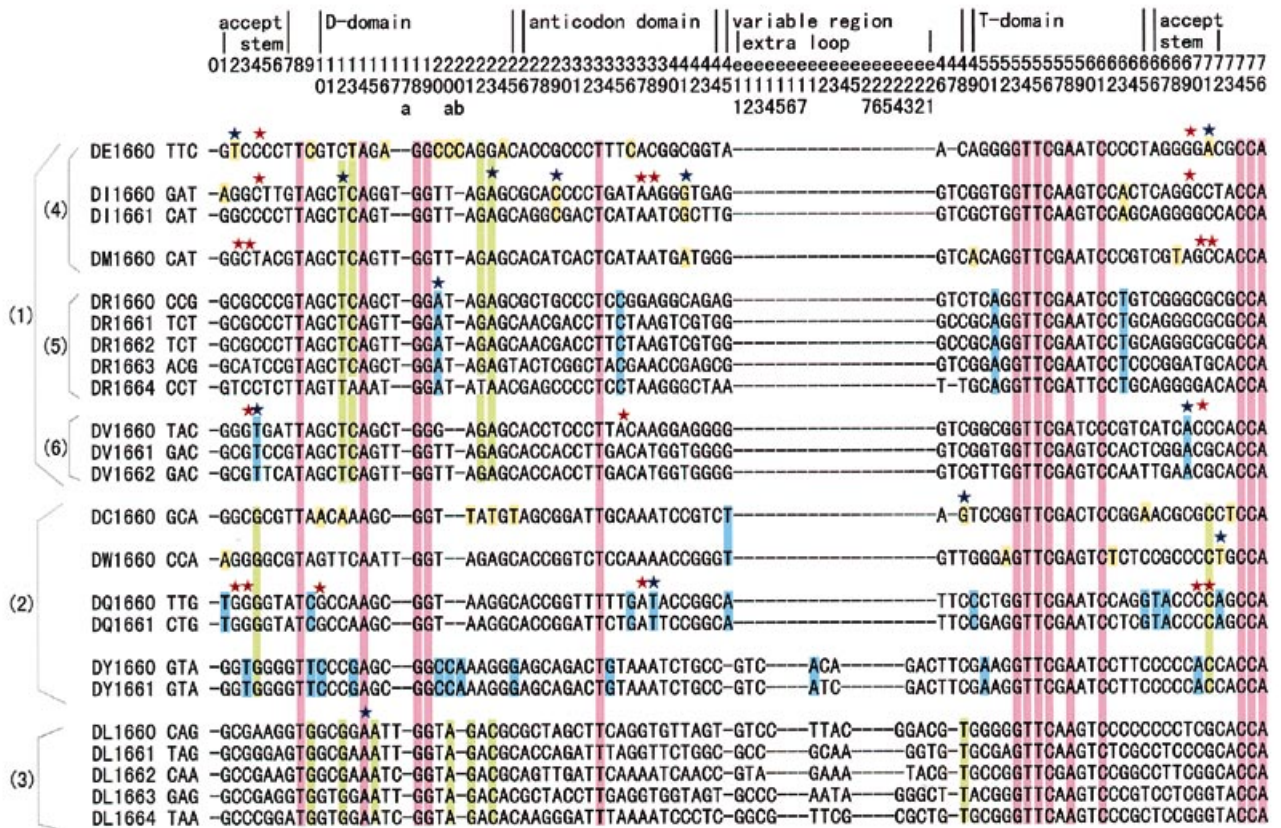


Figure 7. Characteristic bases of each group (Class I). Pink, strictly conserved bases; others, characteristic bases; stars, identities proposed previously.



Figure 8. Characteristic bases of each group (Class II). Pink, strictly conserved bases; others, characteristic bases; stars, identities proposed previously.

are plotted on the sequence space defined by three principal axes corresponding to the three largest eigenvalues, x_1 , x_2 and x_3 . Figure 2B, on the other hand, plots single bases (black circles) in the same space as in Figure 2A. In these figures, distance and direction from the origin have an important meaning in the detection of the characteristic bases: the distance expresses the number of appearances of the base in the position among the all sequences used, whereas the direction from the origin represents the sequence pattern. The first principal axis (x_1) points in the direction of a sequence pattern common to all sequences; bases located in this direction are common to all tRNA sequences used. The second and third principal axes (x_2 and x_3) are group-specific, and we have used them to distinguish the groups of sequences. In Figure 2, sequences are classified into three groups, A, B and C; the groups are characterized by the direction in the x_2 - x_3 plots in Figure 2A, which shows the characteristic sequence pattern of the groups. We can also detect, in Figure 2B, the characteristic bases of the group, which appear in the same direction as the corresponding group. For example, in Figure 2A, characteristic bases of the sequences in group A can be detected in the same direction as the group. The more closely the vector of a base points in the direction of a group, the more specific it is for that group.

RESULTS AND DISCUSSION

In evaluating our method, we used full-length *Escherichia coli* tRNA gene sequences deposited in the aligned sequence database European Bioinformatics Institute (EBI) Data Library (10). We

also used files of the structure of the tRNA^{Gln}-GlnRS crystal that were deposited in the Protein Data Bank (PDB) (11).

Test on tRNA^{Gln}, tRNA^{Leu} and tRNA^{Pro}

We tested our method on three types of tRNAs: tRNA^{Gln}, tRNA^{Leu} and tRNA^{Pro}. Figure 3 shows the plots of the sequences in the two-dimensional plane of the MDS. The x - and y -axes of this figure represent the second and third principal components; the first principal component is not shown in this paper because the direction of its axis represents only the similarity among all the sequences. The tRNA sequences were classified into three groups that correspond to the three types of tRNAs, by using our sequence plots. Figure 4 shows all the bases plotted in the same plane. It is evident that the discriminator (12) of tRNA^{Gln}, 73G, was found among the characteristic bases detected, and that the tRNA^{Leu} and tRNA^{Pro} discriminator, 73A, was found in the direction between tRNA^{Leu} and tRNA^{Pro} directions. This figure shows that we were able to detect bases characteristic to one or more groups of the sequences.

‘Characteristic bases’ are bases that distinguish a sequence group in the total set of sequences considered. In the above test case, base 73A distinguishes the Leu group from the other two groups. However, this base cannot be a characteristic base if we carry out sequence comparison over all types of tRNAs, since 73A also appears in other types of tRNA sequences. The same is true of 73G.



Figure 9. Structure of tRNA^{Gln}-GlnRS complex. Grey, tRNA^{Gln}; blue, GlnRS; pink, 8T 14A 18G 19G 33T 53G 54T 55T 56C 58A 61C 74C 75C 76A; green, 4G 71C; cyan, 9C 36G 38T 45A 49C 65G 66T 67A 72A.

Application to Class I and Class II tRNAs

Since Class I and Class II tRNAs have different tRNA-ARS bindings, we analyzed the two classes separately. Figure 5 shows the result of applying PCA and MDS to Class I tRNAs. As shown in this Figure 5, 23 Class I sequences were classified into three groups: 1, 2 and 3. For this grouping, we took the following strategies: if sequences were of the same species, we classified them into the same group. Otherwise, we grouped the sequences by using the centroid method. Figure 6 shows the result obtained by applying PCA and MDS to group 1 recursively using the same grouping strategies. As shown in this figure, 12 sequences in group 1 were further classified into three groups: 4, 5 and 6. Figures 7 and 8 show the results of comparing the bases of individual sequences for Class I and Class II, respectively. The bases shown in pink represent the bases conserved in all sequences; the bases shown in green represent characteristic bases detected in the first grouping; and the bases shown in blue and yellow represent characteristic bases detected in further grouping of (1) and (2). The bases marked with stars are the identities that have been detected experimentally: blue and red stars, respectively, represent the identities that were and were not also detected by our method.

Approximately 40 and 25% of the identities determined experimentally were detected by our method for Class I and Class II, respectively. In Class I, our method detected many identities in the D-domain. By recursively applying PCA and MDS, our method also detected many experimentally determined identities such as 2T and 71A of tRNA^{Glu}. On the other hand, our method was not able to detect identities G2, G3 and G10 of tRNA^{Gln}, which are known to be important to tRNA^{Gln}-GlnRS recognition (11,13). This recognition is based on the modification (amination)

of G. If we use a base-encoding rule that reflects such modification, we may detect such identities. However, due to the limited number of sequences available, such extension of the encoding rule is expected to introduce further difficulty in the statistical approach.

In Class II, all the experimentally determined identities of tRNA^{Pro} were detected by our method, but the proportion of experimentally determined identities detected by our method in Class II is lower than it is in Class I. The reasons for this might be the following. First, since the experiments for Class II have not been made as intensively as the experiments for Class I analysis, there seems to be many identities that have not been detected in Class II. Secondly, characteristic 'regions' (a region is defined as a series of bases) have not been detected by our method. For example, Phe, Gly and Thr in Class II contain such characteristic regions.

Characteristic bases in the 3D structure of tRNA^{Gln}-GlnRS complex

Most of the experimentally determined identities are in the anticodon arms of tRNAs. McClain *et al.* (8) developed the computer analysis method for detecting identities, and argued that identities detected in the acceptor stem regions are also important to molecular recognition of tRNAs by their cognate ARSs. Our method can also detect characteristic bases in the stem regions in the T and D domains, and these bases are ~56% of the characteristic bases detected by our method. They are located in the elbow region of tRNAs (Fig. 9). For example, G48 of tRNA^{Cys} is an identity that was detected by biochemical experiments (14-18), and the characteristic binding G15-G48 has been considered to be necessary to the tRNA^{Cys}-CysRS recognition. This identity is also detected by our method. The results of the normal mode analysis of tRNAs (19) suggested that the bases in the elbow region play an influential role in determining the dynamics of the molecule. Our result supports this suggestion and we think that the bases in the elbow region determine the structural difference of this region, which in turn affects the interaction between tRNAs and ARSs.

CONCLUSIONS

The method developed here finds identities of tRNA genes by detecting characteristic sites in a group of related sequences. PCA, a basic method of multivariate analysis, is used to classify sequences into groups in a mathematical space. Then, MDS, in which both sequences and bases are represented complementarity in the same mathematical space is used to compare sequences. Analysis of tRNA sequences using MDS has been reported by Nicholas *et al.* (20), but their investigation is limited to sequence grouping. The method we have developed uses not only sequence plots, but also base plots, which enable us to extract the characteristic bases of sequence groups. This is the main advantage of our method. Our method differs from Casari's method (7), which is used to detect functional residues in protein families, in that the classification and comparison procedures are applied recursively in order to classify the sequences into hierarchical groups and detect characteristic sites for each group. This enables us to find subtle patterns of conservation in tRNA sequences.

Approximately 40% of the characteristic sites we detected computationally are identities that have been detected

experimentally, and the other characteristic sites are in the T and D domains (which are in the elbow regions of tRNAs). Thus, it seems that an identity plays a role in determining the structure and dynamics of tRNAs as well as a role in matching tRNAs to their cognate ARSs.

In this work, we used only 23 tRNA gene sequences; thus, the resulting sequence space was rather sparse. The practical advantage of the method becomes apparent when the number of sequences increases. If there are more sequences available, the recursive application can be more effective. The results however show that our method is useful for detecting identities which are difficult to detect by experimentation and other computational methods.

ACKNOWLEDGEMENTS

We thank Professor K.-I.Miura at Gakushuin University, Professors K.Watanabe and T.Ueda at the University of Tokyo, and Mr Bono at Kyoto University for stimulating discussions.

REFERENCES

- 1 Wilcox, M. and Nirenberg, M. (1968) *Proc. Natl. Acad. Sci. USA*, **61**, 229–236.
- 2 Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D. (1990) *Nature*, **347**, 203–206.
- 3 Moras, D. (1992) *Trends Biochem. Sci.*, **17**, 159–164.
- 4 Cavarelli, J. and Moras, D. (1993) *FASEB J.*, **7**, 79–85.
- 5 McClain, W. H. (1993) *FASEB J.*, **7**, 72–78.
- 6 McClain, W. H. (1993) *J. Mol. Biol.*, **234**, 257–280.
- 7 Casari, G., Sander, C. and Valencia, A. (1995) *Nature Struct. Biol.*, **2**, 171–178.
- 8 McClain, W. H. and Nicholas, H. B., Jr (1987) *J. Mol. Biol.*, **194**, 635–642.
- 9 Atilgan, T., Nicholas, H. B., Jr and McClain, W. H. (1986) *Nucleic Acids Res.*, **14**, 375–380.
- 10 Sprintz, M., Steegborn, C., Hübel, F. and Steinberg, S. (1996) *Nucleic Acids Res.*, **24**, 68–72.
- 11 Rould, M., Perona, J. J., Söll, D. and Steitz, T. A. (1989) *Science*, **246**, 1135–1142.
- 12 Crothers, D. M., Seno, T. and Söll, D. G. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 3063–3067.
- 13 Hayase, Y., Jahn, M., Rogers, M. J., Sylvers, L. A., Koizumi, M., Inoue, H., Ohtsuka, E. and Söll, D. G. (1992) *EMBO J.*, **11**, 4159–4165.
- 14 Komatsoulis, G. A. and Abelson, J. (1993) *Biochemistry*, **32**, 7435–7444.
- 15 Pallanck, L., Li, S. and Schulman, L. H. (1992) *Biochemistry*, **267**, 7221–7223.
- 16 Hou, Y.-M., Westhof, E. and Giegé, R. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 6676–6680.
- 17 Hou, Y.-M. (1994) *Biochemistry*, **33**, 4677–4681.
- 18 McClain, W. H. (1993) *J. Biol. Chem.*, **268**, 19398–19402.
- 19 Nakamura, S. and Doi, J. (1994) *Nucleic Acids Res.*, **22**, 514–521.
- 20 Graves, S. B. and Nicjolas, H. B., Jr (1983) *J. Mol. Biol.*, **171**, 111–118.