A Novel Method to Detect *Identities* in tRNA Genes Using Sequence Comparison

Jun-Ichi Sagara [1] jun@bi.a.u-tokyo.ac.jp
Seishi Shimizu [1] seishi@bi.a.u-tokyo.ac.jp
Takeshi Kawabata [2] takawaba@lab.nig.ac.jp
Shugo Nakamura [1] shugo@bi.a.u-tokyo.ac.jp
Mitsunori Ikeguchi [1] ike@bi.a.u-tokyo.ac.jp
Kentaro Shimizu [1] shimizu@bi.a.u-tokyo.ac.jp

[1] Department of Biotechnology, The University of Tokyo
1-1-1 Yayoi, Bunkyo-ku, Tokyo, 113, Japan.

[2] Center for Information Biology, National Institute of Genetics
Yata 1111, Mishima 411, Japan

Abstract

We developed a computational method to detect identities in tRNA genes. The method uses the multidimensional scaling method to classify the sequences of tRNA genes into multiple groups of similar sequences, and also to extract characteristic bases that are conserved within a group but differ from other groups. This procedure was applied recursively to classify the sequences into hierarchical groups so that characteristic sites can be detected more precisely. We were able to detect many characteristic sites in T and D domains of tRNAs as well as the characteristic sites that have been detected experimentally. This suggests that the preservation of L-shape structure in tRNAs is important to the tRNA-ARS recognition.

1 Method

In our method, the multidimensional scaling method (MDS) is firstly applied to the entire sequences of tRNA genes. We define an alignment matrix $\boldsymbol{F}$, each row of which is a sequence vector $\vec{F^k}$ for the kth gene sequence. Each base is encoded to a 4-bit binary number; A, C, G and T are encoded to 1000, 0100, 0010 and 0001, respectively. 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.

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

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 a particular base type exists at a sequence position or not, according to the above encoding rule.

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

$$C^{kk'} = \vec{F^k}\vec{F^{k'}} = \sum_{i,r} F^k_{ir} F^{k'}_{ir} \tag{2}$$

A comparison matrix $\boldsymbol{C}$ $(= C^{kk'})$ with the number of matches for all pairs of sequences can thus be expressed as the matrix product between alignment $\boldsymbol{F}$ and its transpose $\boldsymbol{F}^T$:

$$\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^T \tag{3}$$

The principal axes $\vec{u_p}$ are defined as

$$\boldsymbol{C}\vec{u_p} = \lambda_p \vec{u_p} \tag{4}$$

where $\vec{u_p}$ is an eigenvector and λ_p is the corresponding eigenvalue of comparison matrix $\boldsymbol{C}$. Each sequence is plotted on the two-dimensional plane as shown in Figure 3. The coordinate x_p^k of gene k in dimension p is given by

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

Based on the distance between the two-dimensional sequence plots, sequences are classified into one or more groups.

In our method, bases of each sequence are also projected individually onto the same two-dimensional plane to trace the principal components back to individual bases and positions that characterize individual groups. The coordinates $y_p^{i,r}$ of base r at position i in the sequence are given by

$$y_p^{i,r} = \sqrt{\lambda_p} v_{i,r}^p (= \boldsymbol{F}^T \vec{u}_p)_{ir} \tag{6}$$

We compare the bases with the groups of sequences and detect characteristic bases of each group. The above method is based on Casari's method [1], but we applied this method to tRNA genes [2] while Casari et al. applied this to the Ras-Rab-Rho superfamily. In addition, we applied the above procedure recursively; the groups are classified into subgroups and the characteristic bases can be found in the subgroups. The recursive application of this procedure makes the classification clearer and more unambiguous. This procedure is repeated until the sequences are grouped into single kinds of tRNAs.

2 Summary and Conclusions

We applied the method to Class I tRNAs to detect characteristic sites. We found that about 40% of characteristic sites that we detected are identities that have been detected experimentally, and that the remaining characteristic sites are in T and D domains which are the elbow regions of tRNAs. This result suggests that the characteristic sites in these domains have a role of preserving the L-shape structure in tRNAs.

The practical advantage of the method becomes apparent as the number of sequences increases and as conservation patterns are more subtle. In the application described here, the sequence space is rather sparse; we can use only 23 tRNA gene sequences. If there are more sequences available, the recursive application can be more effective. However, the results show that our method is useful for detecting identities that are difficult to pick out by experimentation and other computational methods.

References

[1] G. Casari and C. Sander and A. Valencia, "A method to predict functional residues in proteins" *Natl. Struc. Biol.*, 2:171-178, 1995.

[2] M. Sprintzl and C. Steegborn and F. Hübel and S. Steinberg "Complsication of tRNA sequences and sequences of tRNA genes" *Nucleic Acids Res.*, 24:68-72, 1996