JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 2, Number 3, 1995
Mary Ann Liebert, Inc.
Pp. 409–416

Analysis of tRNA Gene Sequences by Neural Network

JIAN SUN[1], WEN-YUAN SONG[2], LI-HUANG ZHU[2], and RUN-SHENG CHEN[1]

ABSTRACT

The quantitative similarity among tRNA gene sequences was acquired by analysis with an artificial neural network. The evolutionary relationship derived from our results was consistent with those from other methods. A new sequence was recognized to be a tRNA-like gene by a neural network on the analysis of similarity. All of our results showed the efficiency of the artificial neural network method in the sequence analysis for biological molecules.

Key words: neural network, sequence analysis, similarity, evolution, recognition

INTRODUCTION

AN ARTIFICIAL NEURAL NETWORK is a parallel, distributed information procession system consisting of nonlinear processing elements. In recent years, the surge of research on neural networks gave rise to its extensive application in fields such as associative memory, combinatorial optimization, and pattern recognition.

In 1988, Qian and Sejnowski used the neural network method to predict the secondary structure of global proteins and obtained better results than those from other methods. This shows that as a powerful nonlinear analysis tool, the neural network has its advantage in the application on associative memory. From then on, neural networks were increasingly used in the field of biology.

Since 1989, we have conducted research on the use of neural network for structure prediction and sequence analysis of biomacromolecules. These include (Wang, *et al.,* 1989, 1991; Chen *et al.,* 1990; Sun, 1992; Sun *et al.,* 1991a,b, 1993) two modified ways to train the neural network in prediction of the secondary structure of global proteins, secondary structure prediction of membrane proteins, analysis of the dihedral angles' distribution of proteins, and solution of distance constrains between distant residues in primary sequence of homologous proteins. We also predicted the secondary structure of tRNA, and recognized the splicing sites of pre-mRNA.

There have been many computer methods for identification of the tRNA coding gene. Most of them are based on secondary structure prediction, and the occurrence of invariant or semivariant bases at particular positions, such as Staden (1980), Shortridge *et al.* (1986), and Marvel (1986). Recently, Pavesi *et al.* (1994) described a linear method to search for eukaryotic nuclear tRNA genes in DNA databases; they used a modified weight matrix to recognize two intragenic control regions. Here we use a neural network to study the whole sequences of tRNA genes of viruses, archaebacteria, eubacteria, chloroplasts, mitochondria, and

[1]Institute of Biophysics and [2]Institute of Genetic, Academia Sinica, Beijing 100101, People's Republic of China.

eukaryotes, to compare the sequence similarities among these species and to acquire quantitative relations of similarity. Basing on these similarities, we then recognize a new sequence as a tRNA-like gene sequence.

MATERIALS AND METHODS

The feedforward network is used extensively for structure prediction and sequence analysis in the fields of biology; the algorithm for this network is usually the error backpropagation algorithm (Rumelhart and McClelland, 1986), and details about the application of this network and algorithm can be found in Qian and Sejnowski (1988).

During training, the neural network can build up the correct mapping between the input and the output. Bohr *et al.* (1988) analyzed the homology of proteins by a neural network. Here we use a similar method to analyze similarities among tRNA genes. We take a piece of gene sequence (window) as the input of the neural network, hide the base in the middle position, and train the neural network to predict out this omitted base. By measuring the predictive ability on other sequences, we can compare the similarity between the testing and the trained sequence on the basis of success rate, which is defined as the ratio of the correctly predicted base number to the length of the tested sequence.

According to the size of the trained set, we choose the window length to be 15 bases and the number of the units in the hidden layer is 10, 20, and 60, respectively.

The total sequences used here are all from the tRNA gene sequence compilation by Sprinzl *et al.* (1989). Table 1 gives the constitution of the compilation. According to the number given by Sprinzl, we divide these sequences into the following classes: 0, viruses; 1, archaebacteria; 2, eubacteria; 3, chloroplasts; 4, mitochondria (4–0, single cell organisms and fungi; 4–1, plants; 4–2, animals); 5, single cell organisms and fungi; 6, plants; 7–9, animals (7, *Caenorhabditis elegans, Bombyx mori,* and *Drosophila melanogaster;* 8, *Xenopus laevis* and chicken; 9, mouse, rat, and human).

RESULTS AND DISCUSSION

We use each class of sequences as the training group of the neural network, and the rest of the classes as test groups. Table 2 gives the results of the training and test; in this table each column refers to a network trained by the corresponding class of sequence; the values in it are the test results for the rest of the classes.

TABLE 1. THE SOURCE OF THE tRNA GENES[a]

Name	*Range*	*Sequence*	Base
Viruses	000–099	23	1809
Archaebacteria	100–199	44	3327
Eubacteria	200–299	145	13119
Chloroplasts	300-399	188	14184
Mitochondria	400–499	422	29343
Fungi	400–429	100	6879
Plants	430–449	27	1850
Animals	450–499	295	20614
Eukaryotes	500-999	159	8116
Fungi	500–599	52	3887
Plants	600-699	11	806
Animals	700-999	96	3423

[a]The second column is the range of the index number given by Sprinzl *et al.* (1989), the third column is the number of sequences in the respective class, and the fourth column is the number of bases.

TABLE 2. SUCCESS RATE OF TESTING OF tRNA GENES[a]

	Vir	*Ar*	*Eu*	*Chl*	*Mit*	*Fungi*	*Plants*	*Inv*	*Ver*	*Mam*
Vir	63.8	37.6	45.0	39.5	39.4	35.7	35.4	35.5	33.2	36.4
Ar	35.2	71.2	48.1	44.5	37.7	40.5	39.8	44.8	40.3	42.3
Eu	41.3	42.5	67.0	48.0	41.5	38.2	37.6	40.6	40.2	39.0
Chl	41.1	40.6	48.6	68.7	47.0	38.8	38.6	39.2	38.5	39.3
Mit	31.4	26.5	30.0	32.5	61.4	28.8	27.9	26.5	27.8	27.2
Funti	35.1	38.6	40.3	39.1	40.4	60.2	37.4	42.8	38.7	41.4
Plants	38.1	43.8	45.7	46.5	39.0	44.5	86.7	48.6	46.7	50.6
Inv	33.6	43.0	43.2	40.7	39.6	43.5	42.9	67.9	46.4	52.8
Ver	34.0	41.3	45.6	44.1	41.2	43.7	43.7	52.6	79.8	58.2
Mam	34.3	42.8	42.9	41.3	37.7	44.7	42.4	53.5	51.3	73.0
rand	24.5	26.2	25.9	25.4	28.4	26.5	25.9	27.6	24.3	27.3
short	20.2	28.4	34.8	27.6	24.3	30.1	20.2	30.8	21.4	25.4
long	22.2	27.8	30.3	24.6	26.9	26.7	23.9	27.2	21.8	26.5

[a]Each column gives the success rate of testing for differernt tRNA genes when trained by the respective class in this column. Here the window is 15 bases; the hidden layer has 60 neural units for mitochondria, 20 neural units for eubacteria and chloroplasts, and 10 units for the rest of the classes. In the first column, *rand* is the random sequences obtained by shuffling the respective training group sequences, *long* is the new sequence, and *short* is the sequence that deletes the intron from *long* according to Baldi *et al.* (1992). Vir, viruses; Ar, archaebacteria; Eu, eubacteria; Chl, chloroplasts; Mit, mitochondria; Inv, invertebrates; Ver, vertebrates; Mam, mammals.

From Table 2, we can see that the success rate between any two kinds is almost about 35–55% (in contrast, the random sequences give a success rate of about 25%). This indicates that there are actually similarities and conservation among the tRNA genes. At the same time, we notice that the deviation is always about 4–8%, sometimes even 10%, so there are also differences among classes and within every class. Figure 1 gives the typical results of the success rate for the training group and testing groups, as well as the random sequences. Here we randomly shuffle the sequences in the training group to get the testing random sequences.

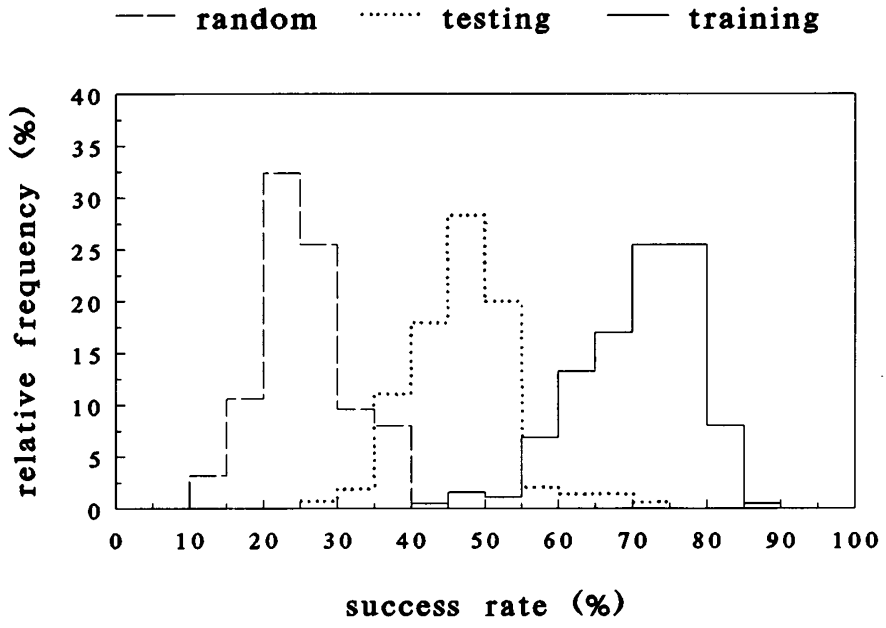

FIG. 1. The typical distribution of success rate on the training class (chloroplasts), testing class (eubacteria), and random sequences. All these are near Gaussian distribution, and there are significant differences between them.

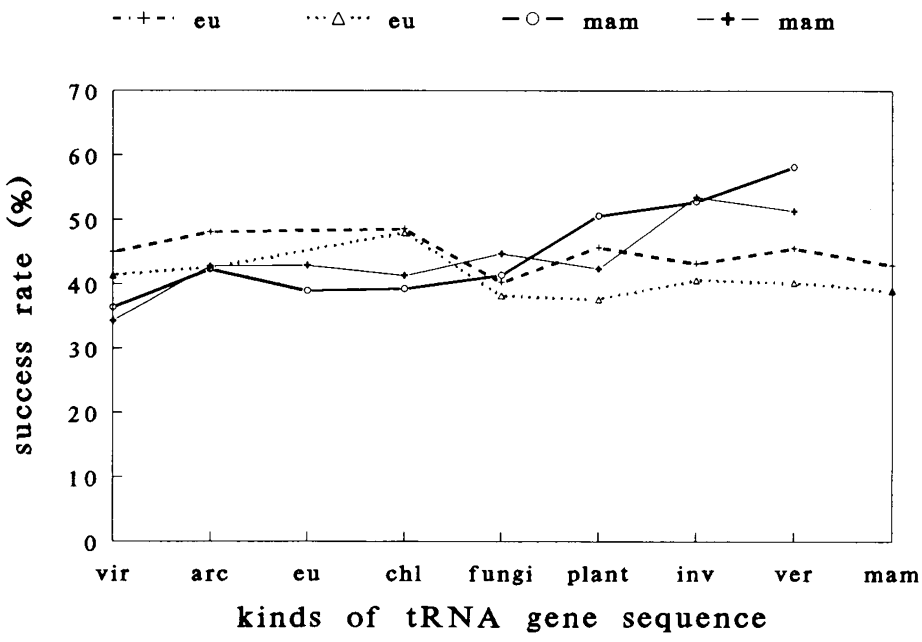

FIG. 2. The similarity of eubacteria and mammals with other classes of sequences. All the data in this figure are from Table 2. The thick lines show the values of the respective column in Table 2, the thin lines are the values of the respective row in Table 2. The figure indicates the differences between prokaryotes and eukaryotes and also shows some linear correlations between species and the similarity; this is obvious in the lines for mammals.

The difference between prokaryotes and eukaryotes

Figure 2 shows the similarity between the tRNA gene sequences of eubacteria (prokaryote) and mammals (eukaryote) with other sequences. In Figure 2, the thick lines indicate the results tested on other sequences while trained on eubacteria or mammals respectively; the thin lines are results tested on eubacteria or mammals while trained on other sequences. We can see that the similarity within prokaryotes is a little higher than that between the prokaryotes and the eukaryotes, and the similarity among the eukaryotes is also higher than that of sequences between eukaryotes and prokaryotes. In Figure 2 the right part of the curve for mammals is obviously higher than the left part. This indicates that there are some differences between tRNA gene sequences of prokaryotes and eukaryotes.

The relation among archaebacteria, eubacteria, and eukaryotes

According to Woese and Fox's (1977) theory of three kingdoms, the prokaryotes are divided into archaebacteria and eubacteria. From Table 2 and Figure 2, we find that the similarity between archaebacteria and eukaryotes is greater than that between eubacteria and eukaryotes. This indicates that archaebacteria are closer to eukaryotes in tRNA gene sequences, which is consistent with the conclusion of Iwabe *et al.* by analyzing the duplicated genes (Iwabe *et al.*, 1989). They concluded that the homology between archaebacteria and eukaryotes is closer than that between archaebacteria and eubacteria, but we cannot verify this from our data.

The origin of chloroplasts and mitochondria

According to the hypothesis of endosymbiosis, since the invasion of prokaryotes into eukaryotes, a stable symbiosis formed; the invader then became an indispensable part of eukaryotes—chloroplasts and mitochondria. As the prokaryotes divided into archaebacteria and eubacteria, we want to determine from which the chloroplasts and mitochondria evolved. From Table 2, we find that chloroplasts and mitochondria are closer to eubacteria than to archaebacteria. Maybe it is the eubacteria from which chloroplasts

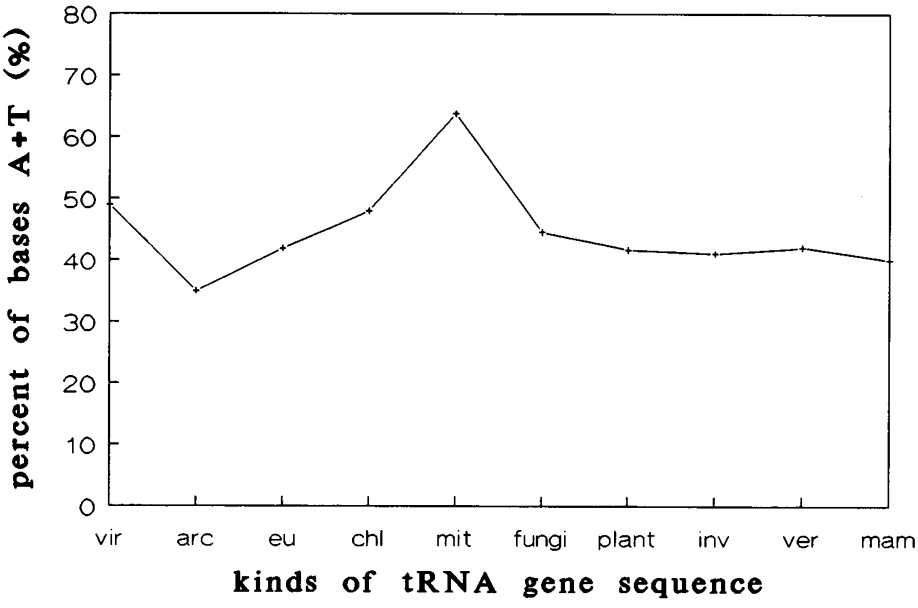

FIG. 3. The percent composition of A+T in every class of tRNA gene. The compositions in all classes are within the range from 35 to 50% except for mitochondria, whose percentage of A+T is 63.8%, much higher than the others. The figure explicitly shows the peculiarity of mitochondria.

and mitochondria evolved. Yang *et al.* (1985) concluded by comparing the sequences of 16 S RNA that the mitochondria and chloroplasts are derived from a group of purple bacteria and its relatives.

The peculiarity of mitochondria

From Table 2 we can see that when mitochondria are tested by other classes, the success rates are very low, almost the same as the random sequences; while trained by mitochondria and testing on other classes, the success rates are within the normal range. This shows that in addition to the common features of tRNA genes, mitochondria tRNA gene sequences have their own peculiarities. Figure 3 shows the base composition of every class of tRNA gene; it shows that the percentage of A+T in mitochondria is over 60%, much higher than that of other classes. The discrepancy of the tRNA genes in mitochondria reflects the extreme diversity of mitochondrial genetic systems.

TABLE 3. SUCCESS RATE OF PREDICTION OF SECONDARY STRUCTURE OF tRNA[a]

	Vir	*Ar*	*Eu*	*Chl*	*Mit*	*Fungi*	*Plants*	*Inv*	*Ver*	*Mam*
Vir	91.9	67.4	76.2	70.8	72.5	67.2	64.7	65.4	62.9	65.7
Ar	71.9	95.2	79.3	76.9	77.0	74.6	75.3	74.8	71.9	76.7
Eu	75.6	74.3	93.9	78.1	75.0	73.3	69.8	71.5	69.1	72.2
Chl	69.8	73.4	80.2	94.3	79.2	74.7	72.3	73.4	70.9	74.2
Mit	64.2	64.5	67.5	67.3	86.8	67.4	63.9	65.4	65.1	64.3
Fungi	67.4	70.0	73.4	74.9	75.5	90.1	69.7	74.1	72.5	76.2
Plants	71.7	75.6	76.9	75.0	77.2	77.3	99.1	80.0	76.2	81.9
Inv	68.1	73.8	78.3	75.6	79.7	80.0	77.7	94.9	79.3	85.0
Ver	68.0	72.3	78.3	75.7	78.9	75.4	77.3	83.6	99.0	86.8
Mam	71.2	74.2	77.1	75.2	78.6	80.4	80.0	84.2	80.2	96.8
Short	57.3	57.2	62.7	60.5	58.8	57.7	70.3	59.8	64.8	62.5
long	51.3	59.6	62.6	62.9	61.8	60.4	67.9	57.9	66.5	62.5

[a]The window is 15 bases and there are 10 neural units in the hidden layer for every class. See footnote to Table 2 for abbreviations.

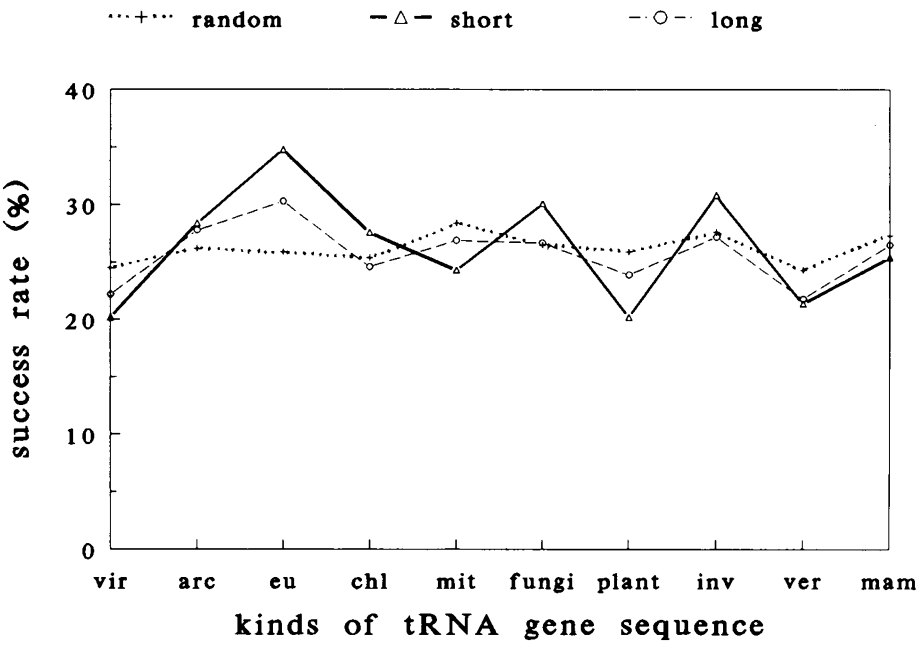

FIG. 4. Comparison of the random sequences with the new sequence. The data are from Table 2. We can see that the new sequence is more similar to the eubacteria tRNA gene and much higher than the random sequences, so we cannot simply regard it as a random sequence.

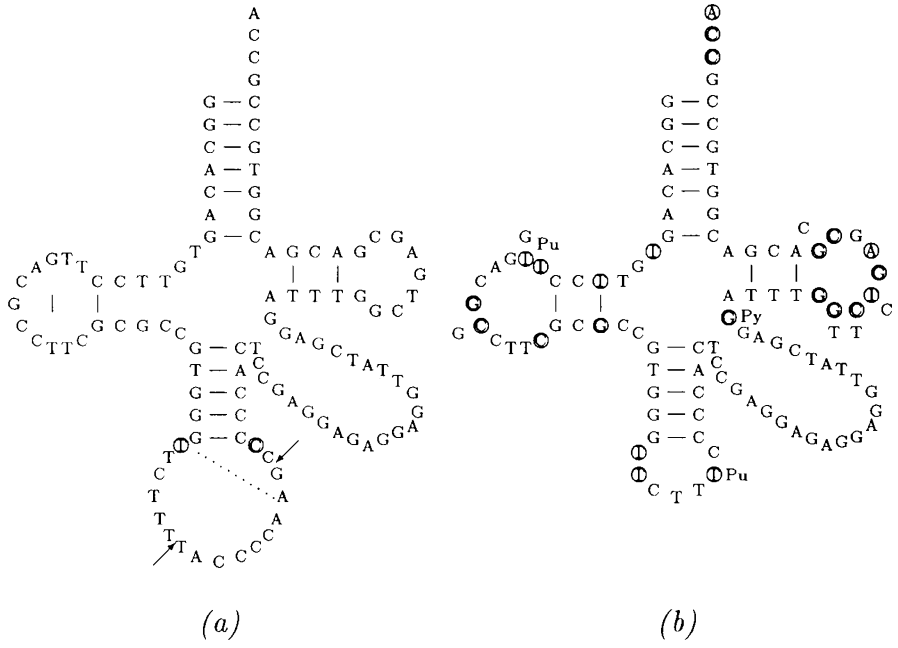

FIG. 5. A tRNA-like structure in the sequence of IR36-L. (a) tRNA-like structure derived from the noncoding strand of nucleotides 100–209 of IR36-L. Arrows indicate the sites for the processing of the anticodon loop–intron according to the rule described by Baldi *et al.* (1992). (b) The tRNA-like structure without the anticodon loop–intron. Circles indicate the conserved base positions. When the circled position is not occupied by the conserved base, the conserved base is given outside the circle.

The analysis of the secondary structure of tRNA genes

We also set up mapping from the primary to the secondary structure by the neural network. According to the alignment of Sprinzl *et al.* (1989), we define the secondary structure of the tRNA gene. The details on the training and testing can be found in Sun (1992). Table 3 gives the results of training and testing. The success rate can also be used as the basis for similarity. By analyzing the results, we can also obtain the above conclusions. At the same time we can see that the secondary structure of tRNA is much more conserved than its primary sequence.

The recognition of tRNA-like gene sequences

From the above we see that although the similarities among all the sequences are different, they always fall into the range of 35–55%. So we can take the success rate of prediction by neural networks as an objective criterion to recognize a tRNA-like sequence. Recently, we found a new sequence in the noncoding strand of the internal transcribed spacer I (ITS1) of the ribosomal RNA gene (rDNA) in rice (Song *et al.*, 1994), which had several characters shared by most tRNA gene sequences. Since tRNA gene sequences have been found in the spacers between 16 S and 23 S rDNA in *E. coli,* chloroplasts, and mitochondria, it is worth searching for the presence of the tRNA gene in rice ITS1. We analyze this sequence by trained neural networks, and give the results in Table 2 and Figure 4. There *long* denotes the new sequence, and *short* denotes the sequence obtained by removing the intron from *long* according to Baldi, *et al.* (1992). The similarity of this new sequence with eubacteria tRNA gene sequences is 34.8% and higher than random sequences 24.9%, and it also has several invariant bases at the conserved sites of the tRNA gene; furthermore we can set up its cloverleaf structure, as shown in Figure 5. We therefore believe this is, indeed, a tRNA-like gene sequence.

CONCLUSION

Macromolecular sequence comparisons are the most accurate and reliable basis to infer phylogenetic relationships. Sequence data are preferable to other molecular methods for assessing evolutionary relatedness because they permit straightforward, quantitative interpretation and, importantly, because they form a growing data base for subsequent reference. Now, as the sequence data on DNA and protein accumulate, it is more important to set up an evolutionary tree with molecular phylogenetics by analyzing these sequences. Our results show that by using a nonlinear tool, the artificial neural network, we can analyze the tRNA gene sequences and obtain results on the evolution that are compatible with other methods applied on the 16 S-like and 23 S-like rRNA sequences. This provides a new method for molecular phylogenetics, and we will continue to do research in this field.

REFERENCES

Baldi, M.R., Mattoccia, E., Bufardeci, E., Fabbri, S., and Tocchini-Valentini, G. 1992. Participation of the intron in the reaction catalyzed by the *Xenopus* tRNA splicing endonuclease. *Science* 255, 1404–1408.

Bohr, N., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Norskov, L., Olsen, O.H., and Petersen, S. 1988. Protein secondary structure and homology by neural networks. *FEBS Lett.* 241, 223–228.

Chen, R.S., Wang, H.J., Shi, X.F., and Ni, X.S. 1990. Predicting the secondary structure of the membrane protein RH and BR using neural network method. *ACTA Biophys. Sinica* 6, 267–270.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. 1989. Evolutioary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic tree of duplicated genes. *Proc. Natl. Acad. Sci. U.S.A.* 86, 9355–9359.

Marvel, C.C. 1986. A program for the identification of tRNA-like structure in DNA sequence data. *Nucleic Acids Res.* 14, 431–435.

Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., and Ottonello, S. 1994. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.* 22, 1247–1256.

Qian, N., and Sejnowski, T.J. 1988. Predicting the secondary structure of globular proteins using neural network methods. *J. Mol. Biol.* 202, 865–884.

Remelhart, D.E., and McClelland, J.L. 1986. *Parrallel Distributed Procession.* MIT Press, Cambridge, MA.

Shortridge, R.D., Pirtel, I.L., and Pirtel, R.M. 1986. IBM microcomputer programs that analyse DNA sequences from tRNA genes. *Comput. Appl. Biosci.* 2, 13–17.

Song, W.Y., Zhang, G.Y., and Zhu, L.H. 1994. A new type of internal transcribed spacer I of ribosomal RNA gene in rice (unpublished data).

Sprinzl, M., Hartmann, T., Weber, J., Blank, J., and Zeidler, R. 1989. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* 17(suppl.), r1–r172.

Staden, R. 1980. A computer program to search for tRNA genes. *Nucleic Acids Res.* 8, 817–825.

Sun, J. 1992. Structure prediction and sequence analysis of biomacromolecules by artificial neural network. Master's thesis, Institute of Biophysics, Academia Sinica.

Sun, J., Ling, L.J., and Chen, R.S. 1991a. Predicting tertiary structure of homologous protein. *High Tech. Comm.* (Chinese) 1(4), 1–4.

Sun, J., Ling, L.J., and Chen, R.S. 1991b. Predicting secondary structure of tRNA through neural network method combined with base pair rule. *High Tech. Comm.*(Chinese) 1(8), 1–4.

Sun, J., Xu, J., Ling, L.J., Shen, R.Q., and Chen, R.S. 1993. Predicting the splicing sites of mRNA by neural network. *ACTA Biophy. Sinica* 9, 127–131.

Wang, H.J., Chen, R.S., Ni, X.S., Shi, X.F., and Ling, L.J. 1989. Neural network methods for predicting the secondary structure of proteins. *ACTA Biophy. Sinica* 5, 422–427.

Wang, H.J., Shi, X.F., Ling, L.J., and Chen, R.S. 1991. Predicting the main-chain dihydral angles of protein using neural network method. *ACTA Biophy. Sinica* 7, 157–160.

Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74, 5088–5090.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G.J., and Woese, C.R. 1985. Mitochondrial origins. *Proc. Natl. Acad. Sci. USA* 82, 4443–4447.

Address reprint requests to:
Run-sheng Chen
Insititute of Biophysics
Academia Sinica
15 Datun Road
Chaoyang District
Beijing 100101, People's Republic of China

Received for publication July 14, 1994; accepted as revised May 15, 1995.