

The Identification on A Small Sample Size of RNA Sequences by Combined Method of Noise Filtering with L_2 -norm

Yu Jen Hu, Yuh-Hua Hu, Jyh Bin Ke, Tin Chi Kuo, Shan Pang Liu, and Ching-Ho Yen

Abstract—This paper proposes a noise filter with L_2 -norm distance method to design a classification of RNA sequences for the species identification, including of the small sample size of the nucleic acid sequence. This method amends and expands the study of Hu et al. in 2011 [1]. We verify this method with the biological sample "slipper orchids" and its hybrid for biological species identification test. The result shows that we can distinguish the paternity of a hybrid among a set of samples of "slipper orchids" by using this method.

Index Terms— L_2 -norm distance; nucleic acid sequence; species identification

I. INTRODUCTION

This method is mainly based on L_2 -norm distance to classify the amino acid sequences, to do pre-processing filtering noise toward the non-A, U, C, G character analyzed through electrophoresis, and to check the progeny of hybrid. This study found that we can easily and efficiently differentiate the species relationships of "slipper orchids" samples by this method, which modified Hu et al.'s study [1]. They explored the sequence analysis but didn't mention the method which might fail. That is, A small sample size of RNA sequence may not be successfully classified by artificial intelligence methods with the mathematical calculation [1] [12]. Pre-processing and noise filtering can solve garbled electrophoresis and effectively resolve the problem of automated RNA sequencing analysis [14] [16] [17]. Consequently, further expansion of the species can truly be applied to biological classification, as Table 2-3.

In the past, the "morphological" observation method [3] was widely adopted to make species identification toward animals and plants. However, the conditions necessary for such identification are very strict. There must be a complete animal and plant appearance or the characteristics parts of that type of animal and plant [2]. RNA records genetic characteristics of organisms, and various species have different genetic composition. Also, different individuals of the same species can be distinguished through RNA analysis.

This study amends the classification of RNA sequences proposed by Hu et al. in 2011 [1], launching mathematical analysis to solve the garbled problem due to the small sample electrophoretic analysis of nucleic acid sequences. RNA electrophoresis analysis has the characteristics of negatively charged nucleic acids which cross the gel in the electric field and move towards the cathode. Because of different

molecular weights, the gel pore size varies in the speed of movement, so as to separate the different sizes of nucleic acids. However, RNA sequencing generally employs vertical electrophoresis [13]. The range of gel electrophoresis analysis can analyze from several nucleotides to millions of chromosomal RNA of nucleotides. However, it has a resolving power within a certain range and can't analyze any RNA fragments of various sizes with a colloid. Therefore, to obtain excellent resolving power, we must explore the range of analytical gel electrophoresis [14].

Two types of gel electrophoresis are commonly used to analyze RNA. One is the agar gel electrophoresis (agarose gel electrophoresis, referred to AGE), the other is polyacrylamide gel electrophoresis (polyacrylamide gel electrophoresis, referred to PAGE) [17]. Because of its concentration in the two different gels, the holes formed in the gels are not the same. Therefore, the scopes of the analysis are different [3].

Today electrophoresis is convenient and reliable to use. However, on the analysis of RNA molecules, it is unable to analyze the chromosome RNA with larger molecules. That is the reason why we need the genes on chromosome localization studies which totally depend on genetic analysis or localization analysis with the microscope [14-17]. This study requires sophisticated artificial experimental operation.

Electrophoresis is caused by nucleic acids in electrophoresis since its own mobile logarithmic rate which is inversely proportional to molecular weight, and it is irrelevant to the base composition and nucleic acid sequences [14] [16]. Nevertheless, there are various causes in the experimental operation and other factors affecting the electrophoresis: (i) colloid concentration, (ii) nucleic acid structure, (iii) electrophoresis buffer salts composition, (iv) electric field strength, (v) electrosmosis phenomenon, (vi) to support the choice of materials, (vii) temperature [14] [16] [17]. Accordingly, it's not easy for us to get a complete noise-free RNA sequence. But, using the appropriate noise filtering pre-processing of this study enables us to resolve the garbled characters in previously mentioned problems and to enhance the accuracy of automated analysis machines.

Through the category of L_2 -norm distance, we achieve the automated species identification with small sample size sequence [4]. With unavailable RNA sequence of trained samples, this study can conduct related calculation of species identification and also supplements how to deal with RNA sequence classification calculations with small samples. It further successfully resolves the issue related to classification, so that future research can take advantage of this principle.

Species identification can be designed to lay the possibility of biological sensors.

Therefore, this study proposes noise filtering pre-processing and L_2 -norm distance for classification. We design a small samples size of RNA sequences (or only single) occurred in the case of classification of biological computing. Also, we use "slipper orchids" to do the actual value of testing biological samples. The results can be found in single RNA hybrid slipper orchids. Some garbled characters in sequence noise filtering can be removed by using pre-processing. Finally, we use L_2 -norm distance classification to distinguish amino acid sequences. The calculation results in this way can be just a small sample of untrained check RNA sequence data. Slipper orchids in this experiment can be found in species identification.

In this study, six native species of "slipper orchids" are inspected and tested in the beginning, then we expand to fourteen native species "slipper orchids" (Source: Council of Agriculture, Executive Yuan, ROC, Taichung District Agricultural Improvement Station) for the fourteen species of slipper orchids native RNA sequence [5]. Then, we calculate a set of hybrid offspring slipper orchid samples. The results show that by employing L_2 -norm distance in the classification, calculated species identification of biological sequence classification can be correctly distinguished, and it further calculates the parent for breeding hybrids of native species and then completes biological calculation of the genetic identification. Consequently, after being tested, this study is considered practical and effective, as shown in table 3-1 to 3-2.

II. MATERIALS AND METHODS

A. Materials

Homogeneous RNA sequence represents having high similarity, coming from the same ancestor, having the same spatial structure, and having similar biochemical functions. Biological definition: if more than 25% of protein amino acid sequence is the same, or more than 75% of the nitrogenous base sequence is the same in RNA, we can conclude that protein or RNA sequence are homogeneous. This point serves as the mathematical calculation reference as we conducted genetic or species identification. Proteins are formed by linear arrangement of amino acid molecules. It is linked through the formation of peptide bonds. Amino acid sequence of the protein is encoded by the corresponding genes. They are mainly 20 standard amino acids encode by the genetic code, as shown in Table 2-1 [7] [8].

Biologists discover the mating phage RNA should be based on the significance of a group of three strings, and it is conducted through the way of Codon. Basically, Codon is the control method of translation when RNA is converted to amino acid sequence. Because there are 20 kinds of amino acids and RNA with 4 bases, RNA is three words as a unit to produce 64 ($4^3=64$) different combinations and it used multivalued function corresponding to 20 amino acids [8].

TABLE 2-1: THE GENETIC CODE TABLE

		Second letter					
		U	C	A	G		
U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U C A G		
	UUC } Phe	UCC } Ser	UAC } Tyr	UGC } Stop			
	UUA } Leu	UCA } Ser	UAA } Stop	UGA } Stop			
	UUG } Leu	UCG } Ser	UAG } Stop	UGG } Trp			
C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U C A G		
	CUC } Leu	CCC } Pro	CAC } His	CGC } Arg			
	CUA } Leu	CCA } Pro	CAA } His	CGA } Arg			
	CUG } Leu	CCG } Pro	CAG } His	CGG } Arg			
A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U C A G		
	AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser			
	AUA } Met	ACA } Thr	AAA } Lys	AGA } Arg			
	AUG } Met	ACG } Thr	AAG } Lys	AGG } Arg			
G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U C A G		
	GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly			
	GUA } Val	GCA } Ala	GAA } Glu	GGA } Gly			
	GUG } Val	GCG } Ala	GAG } Glu	GGG } Gly			

In the genetic code (Table 2-1) shows, Methionine is the general common initiation codon. However, there are very few biological exception is the use of GUG as the initiation codon. UAA, UAG, UGA is the stop codon. They do not correspond to any amino acid, as is the sentence "period". When the translation stop codon when translated if you encounter will stop. Due to base 64 ($4^3=64$) genetic codon, but only 20 kinds of amino acids. Therefore, there must be a lot of duplicate counterparts, such as Arginine is the amino acid corresponding with the most repeated. It can be produced in six different codons.

B. Methods

1) Base sequence noise filtering methods

In this study, in order to address the actual base sequence obtained by electrophoresis of biological samples, it often associated with the experimental data errors occurring phenomena to enhance the computing system the feasibility of automation. For example, it supposed to show AA'A'CCUGGG, but it appeared AA'X'CCUGGG, a garbled problem. Here we designed a new way to solve the noise filter base sequence of occurrence of the above mentioned garbled problems.

The proposed noise filtering is based on electrophoretic analysis of biological experiments [14][15][17]. We take parts of the organizational structure principle when taking the tissue sample, and we divided the above example AAXCCUGGG into two sequences AA + CCUGGG. Because the AA is less than 3 characters, we didn't count them in and only preserved CCUGGG for calculation. We used the genetic code table to translate RNA into protein sequence of the calculation. Finally, Table 2-1 was organized into 22 feature vectors for data analysis as Table 2-2.

TABLE 2-2: 22 AMINO ACID VARIABLE TABLE

Ala (A)	Cys (C)	Asp (D)	Glu (E)	Phe (F)	Gly (G)	His (H)	Ile (I)	Lys (K)	Leu (L)	Met (M)
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Asn (N)	Pro (P)	Gln (Q)	Arg (R)	Ser (S)	Thr (T)	Val (V)	Trp (W)	Tyr (Y)	*	a+t
x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}

$$[x_{1,n}^{(i)} \ x_{2,n}^{(i)} \ \dots \ x_{22,n}^{(i)}], i=1,2,\dots,c \text{ is the number of samples---(1)}$$

We used computer simulation found that classification. If terminator and the words A and T base pairs were been as a paragraph label. There will be 22 parameters. So we let $X_{k,n}^{(i)} = \{x_{1,n}^{(i)}, x_{2,n}^{(i)}, x_{3,n}^{(i)}, \dots, x_{21,n}^{(i)}, x_{22,n}^{(i)}\}$, $x_{k,n}^{(i)}$ represents the k -th characteristic frequency of occurrence in the classification. Then, the number of variables was adjusted. Dimension of the vector was set down to represent the whole sample parameters.

c) Sequence alignment

Tests in this study were calculated by the RNA sequence of the laboratory obtained from Agricultural Improvement area, the biological sequence data. Using the noise filter method of the research conducts sequence data pre-processing. Then we use [8] in the RNA sequence into amino acid sequence principle. Finally, we used our proposed classification of L_2 -norm distance to measure the amino acid sequence existing between the actual gap.

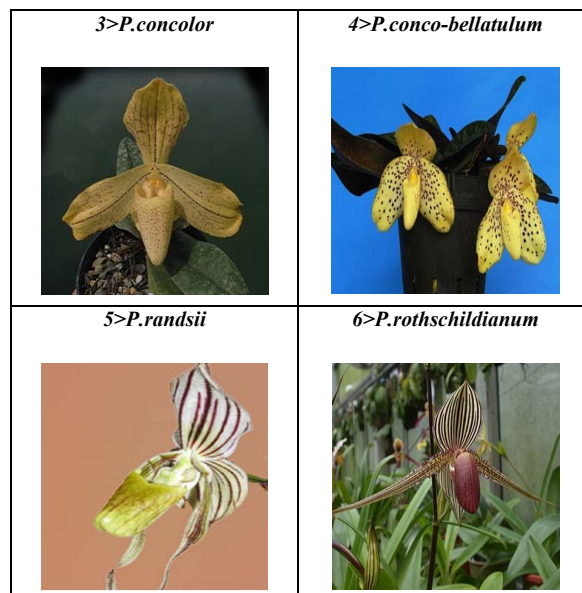
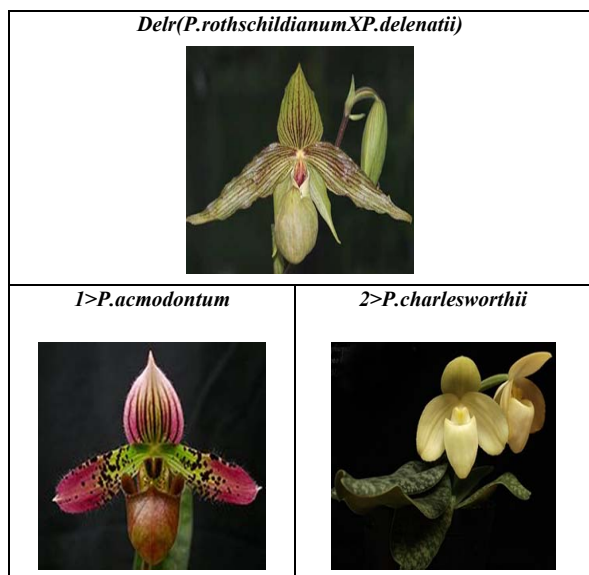
d) Experimental results show that

During the operations in the actual biological experiments, lack of information error is likely to occur. Therefore, we proposed to calculate the noise filter to solve the blind spot. In this study, we used a two-stage biological samples for the actual test, as follows: In the first category, There are six species of slipper orchids, "P.acodontum", "P.charlesworthii", "P.concolor", "P.conco-bellatulum", "P.randsii", "P.rothschildianum", for study samples, and one species, "Delr(P.rothschildianum X P.delenatii)" for the classification of hybrid, and the results are shown in Table 3-1.

TABLE 3-1 : NUMERICAL RESULTS (HYBRIDS):
DELR(P.ROTHSCHILDIANUMXP.DELENATII)

Species	Distance with Delr
<i>P.acodontum</i>	0.007630
<i>P.charlesworthii</i>	0.007152
<i>P.concolor</i>	0.005477
<i>P.conco-bellatulum</i>	0.007368
<i>P.randsii</i>	0.005927
<i>P.rothschildianum</i>	0.003106

TABLE 3-2: TABLE 3-1 PHOTOS.



In the second-staged category, we increased number of the study samples to 14, "P.armeniicum", "P.rothschildianum", "P.chamberlainianum", "P.concolor", "P.glaucophyllum", "P.haynaldianum", "P.lowii", "P.bellatulum", "P.sukhakulii", "P.urbanianum", "P.urbanianum", "P.victoria-mariae", "P.villosum", "P.delenatii", "Phragmipediummem", and the number of hybrids to Magi (P.micranthum X P.delenatii) and use the noise filtering algorithm directly to obtain L_2 -norm distance. The classification result is shown in Table 3-3.

Table 3-3 : NUMERICAL RESULTS (HYBRIDS):












Magi(P.micranthum X P.delenatii)	
Species	Distance with Delr
<i>P.armeniicum</i>	0.003677
<i>P.rothschildianum</i>	0.004269
<i>P.chamberlainianum</i>	0.002896
<i>P.concolor</i>	0.003982
<i>P.glaucophyllum</i>	0.003472
<i>P.haynaldianum</i>	0.005058
<i>P.lowii</i>	0.003744
<i>P.bellatulum</i>	0.004612
<i>P.sukhakulii</i>	0.002515
<i>P.urbanianum</i>	0.002418
<i>P.victoria-mariae</i>	0.003037
<i>P.villosum</i>	0.001740
<i>P.delenatii</i>	0.001378
<i>Phragmipediummem</i>	0.004775





In Table 3-1 to Table 3-3, we could clearly realize the effectiveness and validity of the application in the slipper orchids in this research and know that the minimum L_2 -norm distance on behalf of its parent association or parent.

B. Conclusions

It was common to use the way of diminishing dimension classification forecasts. The advantage of Hu et al. study [1] is that all the dimensions of the sample parameters could be included in the analysis, and more sequence of correct classification out of the group can be found. However, if we encounter the data provided by the native species (parent generation) base sequence and hybrids (offspring) are organized as a single-base sequence, the above approach [1] may be unable to calculate and analyze.

TABLE 3-4: TABLE 3-3 PHOTOS.

<i>Magi(P.micranthum X P.delenatii)</i>	
	
<i>1>P.armeniicum</i>	<i>2>P.rothschildianum</i>
	
<i>3>P.chamberlainianum</i>	<i>4>P.concolor</i>
	
<i>5>P.glaucophyllum</i>	<i>6>P.haynaldianum</i>
	
<i>7>P.lowii</i>	<i>8>P.bellatulum</i>
	
<i>9>P.sukhakulii</i>	<i>10>P.urbanianum</i>
	
<i>11>P.victoria-mariae</i>	<i>12>P.villosum</i>

	
<i>13>P.delenatii</i>	<i>14>Phragmipediummem</i>
	

With the noise filtering, we amended the error produced by the machine through the non-A, U, C, G electrophoresis analysis process. Furthermore, we followed the L_2 -norm distance of the proposed space theory to achieve the species classifications. Finally, we analyzed samples of biological experiments, using native species by the 14 kinds of "slipper orchids" to classify hybrid slipper orchids, and using this research to validate our method in genetic identification and the validity of species identification.

The classification by the numerical results also proved the validity and reasonability of this study. When all the parameters in the classification dimensions are considered, the classification accuracy increases. Additionally, this study proposed noise filtering method and we successfully solved the common biological garbled problem occurred by electrophoresis [14] [17] and completed the error correction. Moreover, we use the actual biological samples of slipper orchids to verify the effectiveness of this method.

This method makes it possible to establish the biological testing simple model of species identification in the future, and makes the automatic detection design more complete and effective.

ACKNOWLEDGMENT

Thank Council of Agriculture, Executive Yuan, Taiwan Associate Professor Y. W. Sun, in this study for their assistance and suggestions.

REFERENCES

- [1] Yu Jen Hu, Yuh Hua Hu, Jyh Bin Ke, The Modified RNA Identification Classification on Fuzzy Relation, Applied Mechanics and Materials Vols. 48-49, pp 1275-1281, 2011; ISSN:17662-7482.
- [2] M. L. Phillips, Crime Scene Genetics: Transforming Forensic Science through Molecular Technologies. BioScience, vol.58, 484-489, 2008; ISSN:0006-3568.
- [3] P. W. Lisette, P. David, Noninvasive Genetic Sampling Tools for Wildlife Biologists: A Review of Applications and Recommendations for Accurate Data Collection, Journal of Wildl. Manage.1419-1433. vol 69,2005; ISSN:1937-2817.
- [4] Xiaohong Wang, Jun Huan, Aaron Smalter, Gerald H Lushington, Application of Kernel Functions for Accurate Similarity Search in Large Chemical Databases, Journal of BMC Bioinformatics, 2010; ISSN:1471-2105.
- [5] Yung Wei Sun, Wen Yi Liao, Han Tsu She, Ming Chung Liu, Yu Ju Liao, Yu Ching Tsai, Chi Hsiung, Junn Jih Chen, Use of Molecular for

Species Identification in Paphiopedilum, Taiwan Flower Expo Flower Posters of New Technology Magazine, 183-186, 2004.

- [6] Chun fen Zhou, Hong wen Peng, Biological Information Easily Learn., Hop Kee Book Press, 2005.
- [7] General Biology-Gene Expression of the Genetic Code, National Yang-Ming University Network Materials.
- [8] Brain Hayes, The Invention of the Genetic Code, American Scientist-Computing Science, Jan.-Feb., 1998; ISSN1545-2786.
- [9] RNA Forensic Science Encyclopedia, R.O.C, Source:<http://www.cib.gov.tw/science/Science0201.aspx?DOC_ID=00007>
- [10] M. Zhang, M. X. Cheng, T. J. Tarn, A Mathematical Formulation of RNA Computation, Journal of IEEE Transaction on Nanobioscience, vol. 5, no.1, 2006; ISSN:1536-1241.
- [11] L. M. Adleman, Molecular Computation of Solutions to Combinatorial Problems, Journal of Science 1021-1024, VOL. 266, 1994; ISSN:0036-8075.
- [12] P. H. William, F. Christophe, G. S. Brian, Fuzzy Species Among Recombinogenic Bacteria, Journal of BMC Bioinformatics, 3:6, 2005; ISSN:1471-2105.
- [13] Summer Basic Molecular Biology Techniques. Genetic Engineering Center, National Chung Hsing University, Taichung, Taiwan, 1999.
- [14] National Pingtung University of Science and Technology, biotechnology, basic experiment, Rui Yu Press, Pingtung, Taiwan, 221, 1998.
- [15] ZENG Yi Xiong, Chen Xinfen, Ching-San Chen, Electrophoretic Separation Symposium, National Science Council, Taipei, Taiwan, 98, 1987.
- [16] Sambrook, J., E. F. Fritsch, and T. Maniatis. Molecular Cloning: a Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. 1989; ISBN-100879695773.
- [17] Li Jianwu et al, Principles and Methods of Biochemical Experiments, Yixuan Book Publishing, 114-146, 2002.



Yu Jen, Hu. Born in Tainan, Taiwan in 1976. Department of Statistics, Feng Chia University, graduated 1998. He graduated from Chung Yuan Christian University, Department of Applied Mathematics in 2000. Ph.D. candidate, National Chung Hsing University. Since 2010 he has been working as math teacher and high school dean in National Experimental High School at Central Taiwan Science Park. Has published papers in:

Magnetic Resonance Imaging, Applied Mechanics and Materials.



Yuh Hua Hu, is an independent researcher. He received his Ph.D. in Computer Science and Information Theory from National Taiwan University, Taiwan. His interest topics are multivariate cryptography, network security and fuzzy.



Jyh Bin, Ke is a Professor of Applied Mathematics at National Chung-Hsing University, Taiwan. He received his MS and Ph. D in Civil Engineering from the University of California, Berkeley, USA. His areas of research include queuing theory, reliability and stochastic modeling. His publications appeared in *Computer and Operational Research, Applied Mathematical Modeling, Mathematical Methods of Operations*

Research, Applied Mathematics and Computation, International Journal of Advanced Manufacturing Technology, Physica A, Journal of Industrial and Management Optimization and others.



Tin-Chi, Kuo is a PhD student of Department of Food Science & Biotechnology, National Chung-Hsing University, Taiwan. Her expertise is in food nutrition, animal experiments and biotechnology. She has published papers in *The Chinese Journal of Physiology*



Shan Pang, Liu, In 2000 he graduated from Graduate Institute of Science Education, National Taiwan Normal University. His master's thesis is related to mathematics learning, curriculum and psychology. In addition, he majored in statistics, and in 2009 received Master Degree in Department of Mathematics, National Taiwan University. His research areas are test theory, reliability and latent trait model. Since 2003 he has been working as high school math teacher. Now he serves at Taipei Municipal Da-Zhi Senior High School, Taipei City, Taiwan.



Ching Ho Yen, is an Assistant Professor of Department of Industrial Engineering and Management Information at Huafan University, where he lectures on operation research, design of experiment and applied statistical analysis. His interests are process capability analysis, acceptance sampling plan and applied statistics. He has published papers in *Technometrics, International Journal of Production Research, Journal of Applied Statistics, Journal of Statistical Computation and Simulation, Communications in Statistics*