Advanced Engineering Forum *Online: 2011-09-09*
ISSN: 2234-991X, Vol. 1, pp 214-220
doi:10.4028/www.scientific.net/AEF.1.214

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

Yu Jen Hu[1,a], Yuh-Hua Hu[2,b], Jyh Bin Ke[1,c], Tin Chi Kuo[3,d],

Ching Ho Yen[4,e], Shan Pang Liu[5,f]

[1]Department of Applied Mathematics, National Chung-Hsing University, Taichung City, Taiwan, R.O.C.
[2]Independent Researcher, Taoyuan County, Taiwan, R.O.C.
[3]Department of Food Science & Biotechnology, National Chung-Hsing University, Taichung City, Taiwan, R.O.C.
[4]Department of Industrial Engineering & Management Information Huafan University, New Taipei City, Taiwan R.O.C
[5]Independent Researcher, Taipei Municipal Da-Zhi Senior High School, Taipei City, Taiwan, R.O.C
[a]herbort_hu@hotmail.com, [b]yuhhuahu@gmail.com, [c]jbke@amath.nchu.edu.tw,
[d]yypp23@hotmail.com, [e]jimyen@cc.hfu.edu.tw, [f]lsp0902@ms54.hinet.net

Keywords: L2-norm distance; nucleic acid sequence; species identification

Abstract. This paper proposed a noise filter with L_2-norm distance method to design a classification of RNA sequences for the species identification, included of the small sample size of the nucleic acid sequence. This method amended and expanded the study by Hu et al. in 2011 [1]. We verified this method with the biological sample "slipper orchids" and its hybrid for biological species identification test. The result is showed that after applied our method, we can distinguish the paternity of a hybrid among a set of samples of "slipper orchids".

Introduction

This method 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 using L_2-norm distance can easily and efficiently differentiate the species relationships of "slipper orchids" samples, which modified Hu et al.'s study [1] which explored the sequence analysis but failed to mention the problem which might exist. That is, with the small sample sized RNA sequence, artificial intelligence methods may not be successfully classified by 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 is 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 different 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 resulted from the small sample electrophoretic analysis of nucleic acid sequences. RNA electrophoresis analysis has the characteristics of negatively charged nucleic acids which will cross the gel in the electric field and move towards the cathode. Because of different molecular weight nucleic acid, 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 gel electrophoresis analysis

capabilities can analyze from several nucleotide to millions of chromosomal RNA of nucleotides. However, it has resolving power within a certain range, not a colloid can analyze any RNA fragments of various sizes. 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 gel, gels formed by the holes 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 the genes on chromosome localization studies totally depend on genetic analysis or localization analysis with the microscope in recent years [14,15,16,17] and it requires sophisticated artificial experimental operation.

Electrophoresis is caused by nucleic acids in electrophoresis since its own mobile logarithmic rate and inversely proportional to molecular weight, and it's not related 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 achieved the automation possibility of species identification with small sample size sequence [4]. With unavailable RNA sequence of training samples in the number of samples, this study could conduct related calculation of species identification and also supplements how to deal with RNA sequence classification calculations with small samples. It further successfully resolved the issue related to classification so that future research can take advantage of this principle. Species identification designed to lay the possibility of biological sensors.

Therefore, this study proposed noise filtering pre-processing and L_2-norm distance for classification. We designed a small samples size of RNA sequences (or only single) occurred in the case of classification of biological computing. Also, we used "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 could be removed by using pre-processing. Finally, we used L_2-norm distance classification to classify 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, the six native species of "slipper orchids" were inspected and tested in the beginning, then we expanded 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]. We calculated a set of hybrid offspring slipper orchid samples. The results are found that by employing L_2-norm distance in the classification, calculated species identification of biological sequence classification could be correctly completed, and it further calculated the parent for breeding hybrids of native species and then completed biological calculation of the genetic identification. Consequently, after being tested, this study could be considered practical and effective, as shown in table 4-1 to 4-7.

Materials and Methods

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

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

In the genetic code (Table 2-1) shows, Methionine is the general common initiation codon 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.

Methods

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}

Table 2-3: 14 native &1 hybrid species of slipper orchids RNA sequence

U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C

Since these 4 characters, A, U, C, G, forms 64 different 3-character strings. From [8], we know what constitute the 20 amino acids of biological proteins. Hence, we adopted the method of Hu et al. [1] and set the codes of 20 kinds of amino acids as parameters $x_1 \sim x_{20}$. Parameters are UAA, UAG, UGA and other "STOP" string as a characteristic frequency of occurrence. Because there is no sequence for the encoded protein fragments, A and U particularly contain more. Therefore, it could be a feature to see if it will be particularly rich in A and U. Therefore, the parameter x_{22} = 『frequency of occurrence of A + U』, as Table 2-2.

Experimental procedures

1. Extracting RNA sequences from slipper orchids: In this study, we mainly extract the ribosomal RNA sequences of ITS1 to ITS2 of slipper orchids. Because this species closely related species RNA sequence. It rapidly evolves and has genetic variation characteristics [5]. Therefore, in the RNA analysis, biologists often use this RNA data.

2. RNA copy (PCR polymerase chain reaction): PCR thermal cycling, RNA sequences are partly copied. Per cycle, twice the number of the original RNA sequence can be obtained. By 30 cycles, we will have two of the thirty times square, that is 1,073,741,824 times, which is about a billion times[7].

3. RNA sequenced (electrophoresis analysis): Sequenced RNA electrophoresis procedure uses PCR analysis. We put the product after the PCR reaction into the automated sequencer for sequence analysis, also adopting the electrophoresis method as our principle. But in the end, we put the laser scanner to scan the base sequence with fluorescent markers, and then determined the RNA sequence we need via the computer [7].

4. The RNA sequence were transformed into the amino acid sequence data and quantified 22 characteristic documents. The research data set is obtained from R.O.C. Council of Agriculture, Taichung District Agricultural Improvement Station provided slipper orchids sequence. Therefore, the starting point of the original sequence is known. There are also some non-A, U, C, G character generator. Therefore, the proposed noise filtering methods were used to fix the garbled problem generated along with the sampling error of the machine.

Results and Conclusions

Results

Sequence analysis of slipper orchids noise filtering

As conducting biological experiments, we found that there were some wrong characters. From the perspective of mathematical analysis and through the discussion on error of the experimental analysis, we did not calculate those wrong characters corresponding to the amino acid variables. If there appeared wrong characters, we analyze data through the algorithm, so that the results of this study could undertake automated calculations. For example: […,..,..,..,AUU,NAC,GCA,..,..,..,…], Because the character, N, is a wrong one, NAC could not be converted into amino acids variable. Therefore we skipped it and did not take NAC into consideration. The longer the whole sequence is, the smaller the error ratio is, as in this formula $\lim_{n\to\infty}\frac{1}{n-1}=0$, the frequency for other characteristics to appear is $\frac{x}{n-1}\fallingdotseq\frac{x}{n}$, and RNA sequence has a certain length, (.. represents a three-character amino acid variables).

Classification of L_2-norm distance

In order to effectively achieve species identification, we design our feature vector $X=\{x_1,x_2,...,x_{22}\}$ based on the amino acids variables in Table 2-1. Then, we set a group of feature vector set toward identified and compared objects, and the feature vector set itself was classified by L_2-norm distance computation in order to reflect the classification of the most essential features (i.e. minimum L_2-norm distance). This is our proposed classification of L_2-norm distance computation process. This program can successfully resolve: To determine the base sequence from the collating sequence with the small sample size and alignment problem, as shown in Table 2-3.

First of all, we transformed bases into 20 groups of amino acids in proteins. An additional group of amino acid bases of gene transfer terminator, and the word A and U base pairs group, as shown in Table 2-2. Amino acids variable served as 22 feature vectors of the study.

We calculated the appearing relative frequency as a feature extraction purposes and converted the original string of data for analysis. The frequency of the first 20 amino acids with terminator genes and A + U base pairs group converted into the amino acid sequence analysis by noise filtering methods to vector representation (1), by the relative frequency of conversion. Then we converted (1) that value into 22 groups. Then we used the smallest sort ($\min\| x_k^i - x_l^j \|$), $i\neq j$, $k\neq l$, $i=1,2,...,c$, $k=1,2,...,n$, $j=1,2...,c$, $l=1,2,...,n$. To find the filter after the first 22 groups of parameters best affinity. Plus terminator and the words A and T base pairs group. Determine the best variables, as (1).

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

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 $X_{k,n}^{(i)}=\{x_{1,n}^{(i)},x_{2,n}^{(i)},x_{3,n}^{(i)},\cdots,x_{21,n}^{(i)},x_{22,n}^{(i)}\}$, $x_{k,n}^{(i)}$ 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.

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.

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.acmodontum***", "***P.charlesworthii***", "***P.concolo***r", "***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 4-1.

Table 4-1：Numerical results (hybrids): *Delr(**P.rothschildianumXP.delenatii**)*

Species	Distance with *Delr*	Species	Distance with *Delr*
P.acmodontum	0.007630	***P.conco-bellatulum***	0.007368
P.charlesworthii	0.007152	***P.randsii***	0.005927
P.concolor	0.005477	***P.rothschildianum***	**0.003106**

In the second-staged category, we increased number of the study samples to 14, "***P.armeniacum***", "***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 4-2.

Table 4-2：Numerical results (hybrids): *Magi(**P.micranthum X P.delenatii**)*

Species	Distance with *Magi*	Species	Distance with *Magi*
P.armeniacum	0.003677	***P.bellatulum***	0.004612
P.rothschildianum	0.004269	***P.sukhakulii***	0.002515
P.chamberlainianum	0.002896	***P.urbanianum***	0.002418
P.concolor	0.003982	***P.victoria-mariae***	0.003037
P.glaucophyllum	0.003472	***P.villosum***	0.001740
P.haynaldianum	0.005058	***P.delenatii***	**0.001378**
P.lowii	0.003744	***Phragmipediummem***	0.004775

In Table 4-1 to Table 4-2, 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.

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.

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.

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.
[2] M. L. Phillips, Crime Scene Genetics: Transforming Forensic Science through Molecular Technologies. BioScience, vol.58, 484-489, 2008.
[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.
[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.
[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, p183-186, 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.
[9] RNA Forensic Science Encyclopedia, R.O.C, 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.
[11] L. M. Adleman, Molecular Compution of Solutions to Combinatorial Problems, Journal of Science 1021-1024, VOL. 266, 1994.
[12] P. H. William, F. Christophe, G. S. Brian, Fuzzy Species Among Recombinogenic Bacteria, Journal of BMC Bioinformatics, 3:6, 2005.
[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, P221, 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.
[17] Li Jianwu et al, principles and methods of biochemical experiments, Yixuan Book Publishing, P114-146, 2002.