ELSEVIER

Computers in Biology and Medicine 38 (2008) 886–893

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Incorporating PCA and fuzzy-ART techniques into achieve organism classification based on codon usage consideration

Kun-Lin Hsieh[a,*], I-Ching Yang[b]

[a] *Department of Information Management, National Taitung University, 684, Sec. 1, Chung - Hua Road, Taitung, Taiwan*

[b] *Department of Natural Science Education, National Taitung University, 684, Sec. 1, Chung-Hua Road, Taitung, Taiwan*

Received 6 November 2006; accepted 19 May 2008

Abstract

To recognize the DNA sequence and mine the hidden information to achieve the classification of organisms are viewed as a difficult work to biologists. As we know, the amino acids are the basic elements to construct DNA. Hence, if the codon usage of amino acids can be analyzed well, the useful information about classification of organisms may be obtained. However, if we choose too many amino acids to perform the clustering analysis, the high dimensions also lead the clustering analysis to be a complicated structure. Hence, in this study, we will incorporate the principle component analysis and fuzzy-ART clustering techniques into constructing an integrated approach. The useful information about organisms classification based on the codon usage can be mined by using the proposed approach. Finally, we also employ a case including 18 bacteria to demonstrate the rationality and feasibility of our proposed approach.

Keywords: Principle component analysis (PCA); Fuzzy adaptive resonance theory (fuzzy-ART); Codon usage

1. Introduction

The classification of organisms can make us to understand the origin of lives. Until now, many studies and researches had focused to address such issue [1–8]. As we know, the coding structure of DNA sequence was frequently used to discuss or to study since meeting the issue of classifying different organisms. Hence, the similarity analysis or clustering analysis of the DNA sequence will be a worthy study to address the classification problem. However, most techniques with the quantitative characteristics cannot be directly employed to DNA sequence. Restated, the transformation for DNA sequence will be necessary action for the subsequent analysis. According to the philosophy of organism evolution, the useful messages can be transferred from DNA to mRNA, and then it will be transferred from mRNA to protein. Next, such useful messages can also be transferred from mRNA to protein via codon types. Generally, each amino acid can match to codon-one at least or it can match to codon-six at most. The codon to encode the same amino acid will be called as the synonymity codon. The frequency of using synonymity codon during the encoding process of protein may be different [9–13], and the particular organism or gene generally focus on one or several specific synonymity codon. Although the codon will be recognized to be a complicated case, it still hidden the important meanings [14,15], e.g. the information providing the recommendation to the classification problem.

In this study, we initially intend to transfer DNA sequence into a quantitative structure based on codon usage. Then, we will apply the transferred form into performing the subsequent clustering analysis. As we know, if we choose too many amino acids to make analysis, it will lead the clustering analysis to meet the case with the multip dimensions [16]. It will cause the clustering analysis to be a complicated operation. Hence, we will also intend to combine the techniques with the dimension reduction characteristic. Hence, we will propose an integrated approach based on soft computing concept, which will incorporate the dimension reduction and clustering technique to resolve the organism classification. Finally, an illustrative data including 18 bacteria will be applied to demonstrating the rationality and feasibility of our proposed approach.

* Tel.: +886 89 318855x1250; fax: +886 89 321981.
E-mail address: klhsieh2644@mail2000.com.tw (K.-L. Hsieh).

0010-4825/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2008.05.007

2. Background review

2.1. Principle component analysis (PCA)

The philosophy of principle component analysis (PCA) can be denoted as that summarizing all parameters to make the necessary analysis since facing such problem [16–19]. PCA will be frequently viewed as one technique to reduce the dimension of problem. Restated, the practitioner can apply PCA to transfer those parameters with the high correlation into the few independent parameters (or it will be called as the principle component term) and the variation of the original data can be still explained well. Those few principle component terms will be the index to explain the summarization of parameters. The equation of PCA can be given as follows:

$$\begin{aligned}\mathrm{PC}(1) &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p\\ \mathrm{PC}(2) &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p\\ &\vdots\\ \mathrm{PC}(m) &= a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p\end{aligned} \tag{1}$$

where $\mathrm{PC}(1), \mathrm{PC}(2), \ldots, \mathrm{PC}(m)$ will denote the first principle component term, the second principle component term, ..., the n-th principle component term. The summarization characteristics will be represented according to the coefficients $a_{11}, \ldots, a_{1p}$ in the linear equation. As for the philosophy to determine those principle terms will be recommended as "The capability of variation explanation for the first principle component term will be the largest one and the capability of the remainder variation for the second principle component term will be the largest one." If we take such philosophy into performing analysis, we will get m $(m \leqslant p)$ principle component terms and the generalized form will be given as follows:

$$\mathrm{PC}(m) = am1X1 + am2X2 + \cdots + ampXp \tag{2}$$

where X_j, $j = 1, 2, \ldots, P$, and we can re-write it into the following equation:

$$Y = \beta 1X1 + \beta 2X2 + \cdots + \beta pXp \tag{3}$$

2.2. Fuzzy adaptive resonance theory (fuzzy-ART)

The behavior in fuzzy adaptive resonance theory (fuzzy ART) lends itself well to simple geometrical interpretation owing to an internal representation of category prototypes as hyperrectangles in the input space. As for the category choice process, by which fuzzy ART always responds the same way to a familiar input: it recalls the smallest hyperrectangle containing this input [20]. Hyperrectangle overlaps have been argued to be an inconvenience if categories are mutually exclusive [21]. In order to learn intersecting and overlapping categories, a neural network must be capable of repressing previously known categories while it forms new ones. In other words, it must be able to make temporary abstraction of previous knowledge. The generalization would allow the learning of the tulips category first, and the flowers category next, whereas discrimination would allow the reverse. In the case of fuzzy ART, increasing the value of a network parameter called vigilance allows formation of new, more specific categories intersecting broad ones that are already known. The network is thus capable of discrimination. However, reducing the same parameter value does not yield generalization. This is due to the predilection of fuzzy ART for the smallest hyperrectangle containing the input. To avoid a category proliferation problem that could otherwise occur [22,23] recommend input normalization by a procedure called complement coding. Let a be an M-dimensional vector $(a_1, a_2, \ldots, a_M)$, where $0 \leqslant a_i \leqslant 1$. The complement coded input I is obtained as $I = (a_1, a_2, \ldots, a_M, 1 - a_1, 1 - a_2, \ldots, 1 - a_M) = (a, a_c)$. Assign to each category j a vector $w_j = (w_{j1}, w_{j2}, \ldots, w_{j2M})$ of adaptive weights. Each category is initially uncommitted, and its weights are initialized to one. The functionality of fuzzy ART may be described as a three-step algorithm [24]:

Step 1. *Category choice*: Upon presentation of an input I, a choice function T_j is computed for each category j.

Step 1. The norm operator $|\bullet|$ is defined as $|x| = \sum_{i=1}^{2M} |x_i|$, the symbol ^ denotes the fuzzy AND operator, that is, $x^\wedge y = (\min(x_1, y_1), \ldots, \min(x_{2M}, y_{2M}))$, and α is a user-defined parameter, $\alpha > 0$. The category J for which the choice function is maximal, that is, $T_j{=}\max\{T_j,\ j{=}1, 2, 3, \ldots\}$ is chosen for the vigilance test.

Step 2. *Vigilance test*: The similarity between w_J and I is compared to a parameter ρ called vigilance, $0 \leqslant \rho \leqslant 1$, in the following test:

$$T_j = \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{\alpha + |\boldsymbol{w}_j|} \tag{1}$$

If the test is passed, then resonance occurs and learning takes place. If the test is failed, then mismatch reset occurs: the value of T_j is set to -1 for the duration of the current input presentation, another category is chosen in Step 1, and the vigilance test is repeated. Categories are searched, that is, chosen and then tested, until one that meets (1) is found. This category is said to be selected for I. It is either already committed or uncommitted, in which case it becomes committed during resonance.

Step 3: *Resonance*: Resonance makes reference to the internal dynamics of the neural network as it pays attention to the vector $(I^\wedge w_J)$. During resonance, the weight vector w_J of the selected category I is updated according to the equation

$$w_J^{(\text{new})} = \beta(I^\wedge w_J^{(\text{old})}) + (1 - \beta)w_J^{(\text{old})} \tag{2}$$

where β is a learning rate parameter, $0 \leqslant \beta \leqslant 1$. What is learned is not the input I itself, but rather an attended weight vector $(I \wedge w_J^{(\text{old})})$: fuzzy ART thus learns prototypes, rather than exemplars. The special case $\beta = 1$ is called fast learning and is assumed throughout this work. Once resonance is finished, a new input may be presented, and the three steps repeated.

3. The proposed approach

In this section, we propose an integrated procedure based on PCA and fuzzy-ART techniques to address the organism classification issue (the flowchart can be graphically depicted in Fig. 1) and it will be summarized as follows.

3.1. Step 1: Preprocessing for DNA sequence

From the logistic concept, DNA will consist of 20 amino acids and the codon type of each amino acid may be not the same (e.g. the type of codon-one or codon-six). However, the practitioners frequently choose the type which can provide more information and the larger codon type may be chosen due to they can provide more hidden information, e.g. the LEU, SER and ARG are the codon-six type. The way to compute the frequency or the ratio for each codon type will be a rational and feasible method to denote the hidden information, i.e. the codon usage. And, it will be included into this study.

3.2. Step 2: Dimension reduction

When the codon type will be chosen, the practitioners will meet the issue of multiple dimensions analysis. From the practice consideration, if the analysis can be performed well via the less dimensions, it will be a suitable option to choose the less dimensions. That is, the dimension reduction will be necessary action. After reviewing the related dimension reduction techniques, the PCA will be a suitable technique to address such issue. In order to make the dimension reduction via PCA, a suitable cutoff value (e.g. 95% or 0.95) to the explanation capability can be determined at initial stage. Then, the number of the principle component terms can be determined after performing the PCA. After completing the PCA, we can apply the Eq. (2) to obtain the corresponding principle component values, i.e. the transferred data after dimension reduction.

3.3. Step 3: Making clustering analysis

Next, we will process the clustering analysis according to the obtained principle components. The fuzzy-ART will be applied to find out the possible clustering result. Generally, the try-and-error to the vigilance value (it will lie in 0–1) will be used to aid the decision-making process about the number of clusters. Herein, we will intend to adjust the vigilance value from 0.5 to 0.9. The primary consideration about such choice can be explained as "Too larger vigilance value will stress the higher similarity and too less vigilance value will stress the better plasticity. The baseline of the vigilance can be determined as 0.5 due to that it is the balance point to similarity and plasticity." Then, the number of cluster can be then determined according to the compromise of clustering results derived from the different vigilance values. Firstly, we can list all possible cluster results and compute the frequency depending on the count of appearance for each particular cluster. Next, we will compute the total frequency score (i.e. total frequency score = frequency − 1) for each cluster. The designed reason of total frequency score is to obtain the repetition of such cluster. Then, the number of cluster with the maximum frequency score will be determined as the optimum clustering result. And, the vigilance value can also be determined as the maximum vigilance value under the optimum clustering result. As for the detailed operational procedure for performing fuzzy-ART, it will be referred to Section 2.

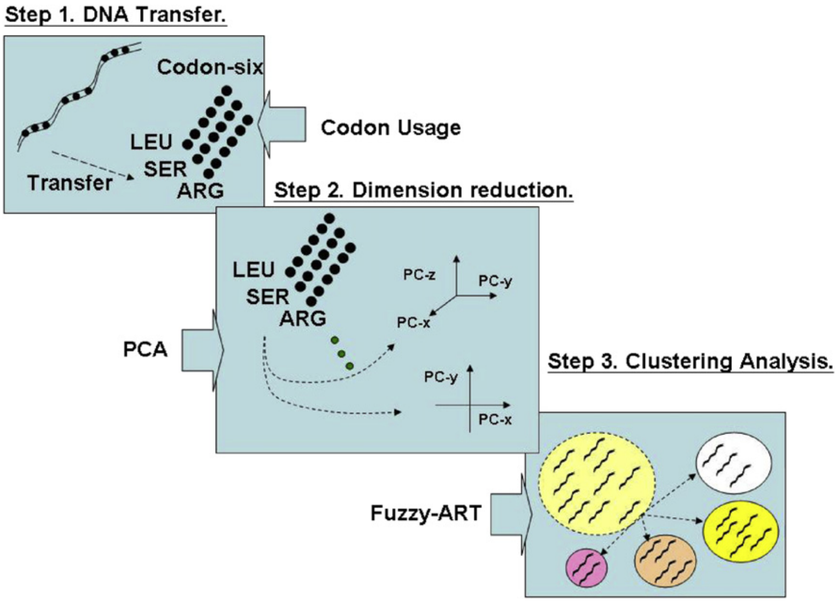


Fig. 1. The flowchart of the proposed procedure.

4. Illustrative example

In this section, we will apply an illustrative case to demonstrate the rationality and feasibility of our proposed approach.

4.1. Materials

We collect a data from Gonome Altas Database (it can be obtained from http://www.cbs.Dtu.dk/services/GenomeAtlas/) to perform the analysis. We download about 18 bacteria with the full DNA sequence data and it will be given in Table 1.

4.2. Analysis procedure

4.2.1. Step 1: Preprocessing for DNA sequence

Herein, three amino acids (LEU, SER, ARG) with codon-six will be chosen to be included in this study. The primary reason is that the codon-six can hold more sufficient information with the simple structure Then, the original DNA sequence will be transferred into the codon usage in this study. The related data can be collected from database and the frequency of codon usage will be computed as Table 2 by referring to [10,14,15]. That is, the DNA sequence of each bacteria can be transferred

Table 1
The related data about the 18 bacteria

No.	Organism	Label	Accession no.	Bases (bp)	Taxo. ID
1	*Bacillus anthracis* str. Ames	B1	AE016879	5227293	198094
2	*Bacillus anthracis* str. 'Ames Ancestor'	B2	AE017334	5227419	261594
3	*Bacillus anthracis* str. Sterne	B3	AE017225	5228663	260799
4	*Bacillus cereus* ATCC 10987	B4	AE017194	5224283	222523
5	*Bacillus cereus* ATCC 14579	B5	AE016877	5411809	226900
6	*Bacillus cereus* ZK	B6	CP000001	5300915	288681
7	*Bacillus clausii* KSM-K16	B7	AP00627	4303871	66692
8	*Bacillus halodurans* C-125	B8	BA000004	4202352	272558
9	*Bacillus licheniformis* ATCC 14580	B9	AE017333	4222645	279010
10	*Bacillus subtilis* subsp. subtilis str. 168	B10	AL009126	4214630	224308
11	*Bacillus thuringiensis* serovar *konkukian* str. 97-27	B11	AE017355	5237682	281309
12	*Escherichia coli* CFT073	E1	AE014075	5231428	199310
13	*Escherichia coli* K12	E2	U00096	4639675	83333
14	*Escherichia coli* O157:H7	E3	BA000007	5498450	83334
15	*Escherichia coli* O157:H7:EDL933	E4	AE005174	5528445	155864
16	*Pyrococcus abyssi* GE5	P1	AL096836	1765118	272844
17	*Pyrococcus furiosus* DSM 3638	P2	AE009950	1908256	186497
18	*Pyrococcus horikoshii* OT3	P3	BA000001	1738505	70601

Table 2
The frequency of the codon usage of LEU, SER and ARG for the 18 bacteria

Codon usage	LEU						SER						ARG					
Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
B_anthracis-Ames	0.521	0.099	0.190	0.044	0.111	0.034	0.266	0.055	0.252	0.079	0.248	0.099	0.377	0.128	0.144	0.036	0.251	0.065
B_anthracis-Ames0581	0.521	0.099	0.190	0.044	0.111	0.034	0.266	0.055	0.253	0.079	0.249	0.099	0.377	0.128	0.144	0.036	0.251	0.065
B_anthracis-Sterne	0.520	0.101	0.189	0.044	0.111	0.034	0.265	0.055	0.252	0.079	0.249	0.099	0.373	0.127	0.145	0.036	0.252	0.066
B_cereus-ATCC10987	0.523	0.105	0.188	0.047	0.111	0.036	0.262	0.058	0.250	0.080	0.247	0.102	0.361	0.124	0.145	0.041	0.257	0.073
B_cereus-ATCC14579	0.518	0.104	0.191	0.042	0.111	0.034	0.265	0.053	0.255	0.079	0.249	0.099	0.373	0.122	0.144	0.036	0.259	0.067
B_cereus-ZK	0.521	0.099	0.189	0.044	0.112	0.034	0.268	0.056	0.252	0.080	0.247	0.098	0.373	0.126	0.146	0.036	0.253	0.066
B_claussi-KSMK16	0.219	0.247	0.228	0.106	0.093	0.107	0.173	0.123	0.184	0.160	0.125	0.235	0.231	0.265	0.125	0.187	0.105	0.087
B_halodurans-C125	0.274	0.174	0.219	0.138	0.115	0.081	0.175	0.142	0.201	0.159	0.153	0.170	0.244	0.179	0.223	0.155	0.127	0.072
B_licheniformis-DSM13	0.143	0.181	0.229	0.153	0.032	0.263	0.147	0.155	0.202	0.170	0.063	0.264	0.115	0.254	0.073	0.193	0.218	0.146
B_B_subtills-168	0.198	0.159	0.239	0.112	0.051	0.240	0.204	0.128	0.236	0.101	0.106	0.226	0.180	0.206	0.100	0.157	0.260	0.096
thuringiensis-9272	0.522	0.100	0.189	0.045	0.111	0.034	0.266	0.055	0.251	0.080	0.249	0.099	0.373	0.129	0.145	0.037	0.250	0.065
E_coli-CFT073	0.132	0.134	0.111	0.104	0.038	0.481	0.146	0.148	0.133	0.159	0.159	0.268	0.262	0.272	0.052	0.084	0.123	0.207
E_coli-K12	0.131	0.128	0.104	0.105	0.037	0.496	0.145	0.149	0.124	0.151	0.151	0.277	0.378	0.398	0.066	0.099	0.038	0.022
E_coli-O157:H7	0.133	0.125	0.110	0.102	0.038	0.493	0.144	0.150	0.136	0.156	0.156	0.268	0.361	0.374	0.069	0.111	0.052	0.033
E_coli-O157:H7-EDL933	0.140	0.134	0.112	0.102	0.043	0.470	0.147	0.147	0.146	0.152	0.152	0.255	0.316	0.343	0.091	0.134	0.066	0.051
P_abyssi-GE5	0.163	0.143	0.208	0.202	0.181	0.104	0.138	0.144	0.180	0.151	0.151	0.281	0.018	0.020	0.015	0.011	0.295	0.640
P_furiosus-DSM3638	0.194	0.139	0.240	0.169	0.176	0.082	0.184	0.127	0.208	0.208	0.208	0.221	0.021	0.021	0.023	0.014	0.524	0.397
P_horikosgii-OT3	0.208	0.142	0.239	0.171	0.151	0.089	0.213	0.205	0.083	0.157	0.157	0.167	0.092	0.053	0.101	0.074	0.328	0.352

Table 3
The statistical test for principle component terms

Total variance explained									
Component	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative (%)	Total	% of variance	Cumulative (%)	Total	% of variance	Cumulative (%)
1	9.399	52.216	52.216	9.399	52.216	52.216	7.865	43.696	43.696
2	5.043	24.015	80.230	5.043	24.015	80.230	5.376	29.869	73.566
3	2.021	11.23	91.459	2.021	11.225	91.459	3.221	17.594	91.459
4	0.654	3.635	95.094						
5	0.363	2.018	97.113						
6	0.267	1.481	96.593						
7	0.170	0.944	99.538						
6	4.484E−02	0.249	99.787						
9	2.902E−02	0.161	99.948						
10	7.420E−05	4.122E−02	99.999						
11	1.389E−05	7.718E−03	99.997						
12	3.976E−04	2.209E−03	99.999						
13	1.372E−04	7.621E−04	100.000						
14	4.116E−06	2.287E−04	100.000						
15	1.775E−06	9.862E−06	100.000						
16	6.406E−06	3.559E−07	100.000						
17	4.318E−11	2399E−10	100.000						
IS	5.763E−16	3.202E−15	100.000						

Extraction method: Principal component analysis.

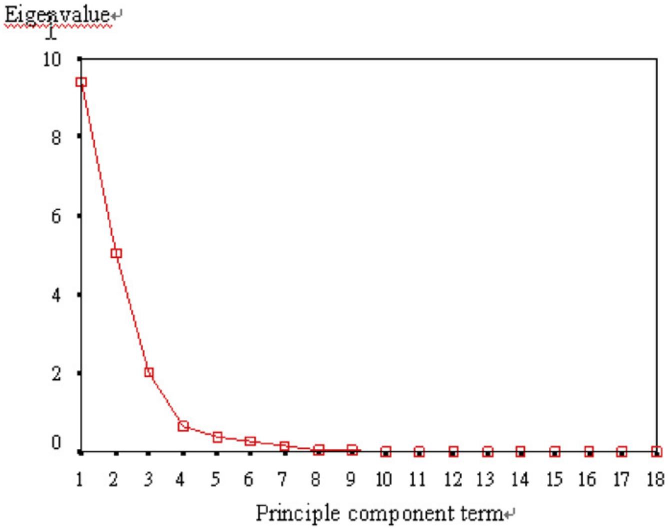
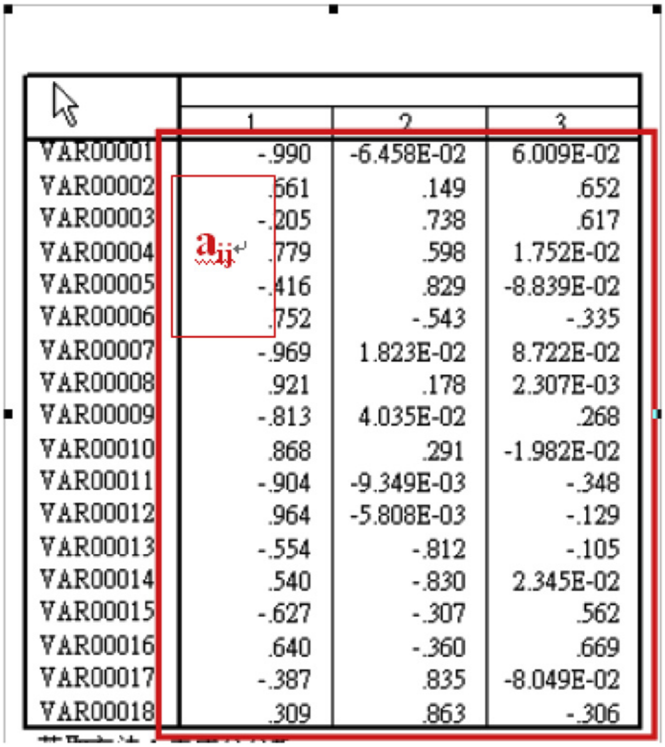


Fig. 2. The scree plot.

into 18 values (or dimensions). That is, it can be viewed as the quantification procedure.

4.2.2. Step 2: Dimension reduction

To simplify the analysis operation, we will apply the software SPSS10.0 to perform the PCA to achieve the dimension reduction action. We initially assign a cutoff value C_v in PCA to be 0.9 with a strict consideration. Restated, the explanation capability of the whole variation for the chosen principle component terms must exceed 90%. Under such constrain, we choose three principle components in this study due to the explanation capability of the whole variation can arrive at about 91.5%. The related statistical report can be listed in Table 3 and the option of principle component terms can be also referred to the scree plot in Fig. 2. Besides, the linear structure for the three chosen principle component terms can refer to Fig. 3. For



	1	2	3
VAR00001	-.990	-6.458E-02	6.009E-02
VAR00002	.661	.149	.652
VAR00003	-.205	.738	.617
VAR00004	.779	.598	1.752E-02
VAR00005	-.416	.829	-8.839E-02
VAR00006	.752	-.543	-.335
VAR00007	-.969	1.823E-02	8.722E-02
VAR00008	.921	.178	2.307E-03
VAR00009	-.813	4.035E-02	.268
VAR00010	.868	.291	-1.982E-02
VAR00011	-.904	-9.349E-03	-.348
VAR00012	.964	-5.808E-03	-.129
VAR00013	-.554	-.812	-.105
VAR00014	.540	-.830	2.345E-02
VAR00015	-.627	-.307	.562
VAR00016	.640	-.360	.669
VAR00017	-.387	.835	-8.049E-02
VAR00018	.309	.863	-.306

Fig. 3. The coefficients in three chosen principle component terms.

Table 4
The comparison result for the chosen topological structure and the final number of cluster

The chosen vigilance value	0.5	0.55	0.6*	0.65
The number of cluster after clustering	4	4	4*	5
The corresponding members	Cluster1: {1, 2, 3, 4, 5, 6, 11}(2) Cluster2: {7, 8, 9, 10, 11}(0) Cluster3: {12, 13, 14, 15}(4) Cluster4: {16, 17, 18}(3)	Cluster1: {1, 2, 3, 4, 5, 6, 11}(2) Cluster2: {7, 8, 9, 10}(2) Cluster3: {12, 13, 14, 15}(4) Cluster4: {16, 17, 18}(3)	Cluster1: {1, 2, 3, 4, 5, 6, 11}(2) Cluster2: {7, 8, 9, 10}(2) Cluster3: {12, 13, 14, 15}(4) Cluster4: {16, 17, 18}(3)	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {7, 8, 9, 10}(2) Cluster3: {13, 14, 15}(0) Cluster4: {12}(0) Cluster5: {16, 17, 18}(3)
Total frequency score	9	11	11	5
The chosen vigilance value	0.7	0.75	0.8	0.85
The number of cluster after clustering	7	9	10	10
The corresponding members	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {7, 8, 9}(0) Cluster3: {12, 13, 14, 15}(4) Cluster4: {10, 11}(0) Cluster5: {16}(0) Cluster6: {17}(0) Cluster7: {18}(0)	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {8, 9}(0) Cluster3: {10}(0) Cluster4: {11}(0) Cluster5: {12, 13, 14, 15}(4) Cluster6: {7}(0) Cluster7: {16}(0) Cluster8: {17}(0) Cluster9: {18}(0)	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {7}(0) Cluster3: {8, 9}(0) Cluster4: {10}(0) Cluster5: {11}(0) Cluster6: {12}(0) Cluster7: {13, 14, 15}(0) Cluster8: {16}(0) Cluster9: {17}(0) Cluster10: {18}(0)	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {7}(0) Cluster3: {8, 9}(0) Cluster4: {10}(0) Cluster5: {11}(0) Cluster6: {12}(0) Cluster7: {13, 14, 15}(0) Cluster8: {16}(0) Cluster9: {17}(0) Cluster10: {18}(0)
Total frequency score	4	4	0	0
The chosen vigilance value	0.9			
The number of cluster after clustering	12			
The corresponding members	Cluster1: {1, 2, 3, 4, 5, 6}(0) Cluster2: {7}(0) Cluster3: {8, 9}(0) Cluster4: {10}(0) {11}(0) Cluster6: {12}(0) Cluster7: {13}(0) Cluster8: {14}(0) Cluster9: {15}(0) Cluster10: {16}(0) Cluster11: {17}(0) Cluster12: {18}(0)			
Frequency score	0			

{ bacteria } (frequency score) will denote the frequency score of bacteria.

*It will denote the optimum number of cluster and the optimum vigilance value.

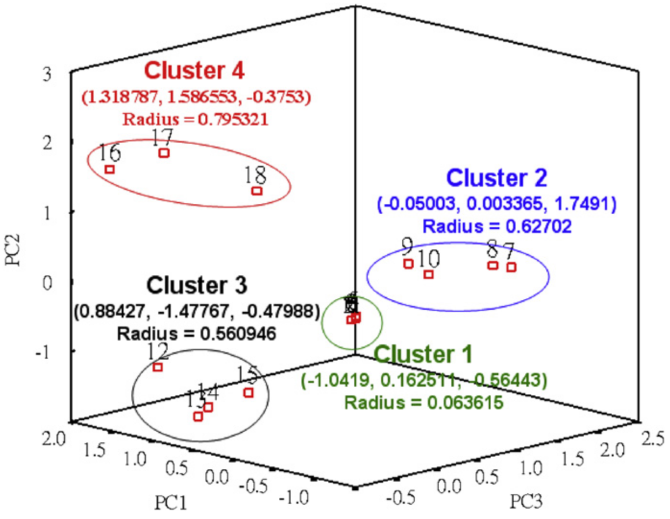


Fig. 4. The 3D clustering diagram for the proposed approach.

instance, the PC(1) can be represented as

$$
\begin{aligned}
\mathrm{PC}(1) = {} & -0.99 * \mathrm{var1} + 0.561 * \mathrm{vqr2} - 0.205 * \mathrm{var3} + \cdots \\
& + 0.640 * \mathrm{var16} - 0.387 * \mathrm{var17} \\
& + 0.309 * \mathrm{var18} \qquad (3)
\end{aligned}
$$

As for PC(2) and PC(3), the linear structure can also be constructed by using the same concept like as PC(1).

4.2.3. Step 3: Making clustering analysis

Subsequently, we will perform clustering analysis by using fuzzy-ART to obtain the possible number of cluster. The possible vigilance value will be taken from 0.5 to 0.9. For simplifying the operation process, we choose the vigilance values with the same increment 0.05. That is, we will perform nine times clustering analysis by using fuzzy-ART. Table 4 will list the clustering result for the chosen vigilance value.

From Table 4, we can determine the feasible number of cluster to be four. The reason about making such decision can be denoted as "The frequency score obtained from the case with four clusters is the maximum (i.e. total frequency score = 11). It denoted the optimum number of cluster can be determined as four and the corresponding vigilance value can be determined as 0.6 (i.e. 0.6 > 0.55)". Besides, we can also graphically depict such result via a 3D visual diagram with the vigilance value equaling to 0.6. The constructed 3D diagram can be represented in Fig. 3. From Fig. 4, not only the clustering effect can be shown, but the radius for such four clusters and the center of cluster also can be represented. The radius can be computed by choosing the maximum distance value for all organisms included in such cluster with the center of cluster. The biologists may make more detailed study after reviewing the 3D clustering diagram.

According to the result of clustering analysis, we can summarize several findings:

(1) *P_abyssi*-GE5 (No. 16), *P_furiosus*-DSM3638 (No. 17) and *P_horikosgii*-OT3 (No. 18) can be recommended to be clustered into the same cluster.

(2) *E_coli*-CFT073 (No. 12), *E_coli*-K12 (No. 13), *E_coli*-O157:H7 (No. 14), *E_coli*-O157:H7-EDL933 (No. 15) can be then recommended to clustered into the same cluster.

(3) As for the other 11 bacteria, from *B_anthracis*-Ames to *B_thuringiensis*-9272, we will suggest that two groups can be clustered. Among those bacteria, No. 1, 2, 3, 4, 5, 6 and 11 will be clustered to the same cluster. This finding has the same result as [24]. From the biological philosophy, those bacteria are the member of *Bacillus cereus*. Besides, it also denotes that *Bacillus anthracis* and *Bacillus thuringiensis* can be resolved to have the direct evolution link to *B. cereus* after the sequence analysis of chromosome. Furthermore, *Bacillus licheniformis* ATCC 14580 (No. 9) and *Bacillus subtilis* subsp. *subtilis* str. 168 (No. 10) had been found out that the two organism will have a higher similarity for the organization. It will be recommended as the same evolution chain. As for *Bacillus clausii* (No. 7) and *Bacillus halodurans* (No. 8), they will be recommended to cluster into the same cluster. Although there is not any direct research to explain such situation, we can make suggestion to the future search.

5. Concluding remarks and recommendations

In this study, we proposed an integrated approach incorporating PCA and fuzzy-ART techniques to achieve the organism's classification from the viewpoint of codon usage. The primary contribution is to provide a rational and feasible integration approach to analyze amino acids. It can be viewed as an optional reference for biologists in the future. Besides, we also apply an illustrative example to demonstrate the proposed approach. The findings almost have the same result or recommendation to the previous researches. As for the initial stage of study or research, the biologists can apply this systematic approach to shorten the preprocessing time and capture the possible reference information. Therefore, the power of the proposed approach can be enhanced via incorporating the other techniques, e.g. the case-based reasoning (CBR), expert systems or decision supporting system (DSS) in the future.

Conflict of interest statement

None declared.

References

[1] Y.H. Chen, S.L. Nyeo, J.P. Yu, Power-laws in the complete sequences of human genome, J. Biol. Syst. 13 (2005) 105–115.

[2] M. De Sousa Vieira, Statistics of DNA sequences: a low-frequency analysis, Physical Review E 60 (1999) 5932–5937.

[3] R.F. Doolittle, D.F. Feng, S. Tsang, G. Cho, E. Little, Determining divergence times of the major Kingdoms of living organisms with a protein clock, Science 271 (1996) 470–477.

[4] Y. Isohata, M. Hayashi, Analyses of DNA base sequences for eukaryotes in terms of power spectrum method, Jpn. J. Appl. Phys. 44 (2005) 1143–1146.

[5] S.L. Nyeo, I.C. Yang, C.H. Wu, Spectral classification of archaeal and bacterial genomes, J. Biol. Syst. 10 (2002) 233–241.

[6] R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805–3808.

[7] M.W. Rey, P. Ramaiya, B.A. Nelson, S.D. Brody-Karpin, E.J. Zaretsky, M. Tang, A.L. de Leon, H. Xiang, V. Gusti, I.G. Clausen, P.B. Olsen, M.D. Rasmussen, J.T. Andersen, P.L. Jørgensen, T.S. Larsen, A. Sorokin, A. Bolotin, A. Lapidus, N. Galleron, S.D. Ehrlich, R.M. Berka, Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species, Genome Biol. 5 (10) (2004) R77.

[8] J.G. Holt, N.R. Krieg, P.H.A. Sneath, J.T. Staley, S.T. Williams, Bergey's Manual of Determinative Bacteriology, ninth ed., Williams & Wilkins, 1994.

[9] T. Ghosh, Studies on codon usage in *Entamoeba histolytica*, Int. J. Parasitol. 30 (2000) 715–722.

[10] S. Karlin, J. Mrazek, What drives condon choices in human genes?, J. Mol. Biol. 262 (1996) 459–472.

[11] M.R. Cancilla, A.J. Hillier, B.E. Davidson, *Lactococcus lactis* glyceraldehyde-3-phosphate dehydrogenase gene, gap—further evidence for strongly biased codon usage in glycolytic pathway genes, Microbiology-UK 141 (1995) 1027–1036.

[12] M.A. Freirepicos, M.I. Gonzalezsiso, E. Rodriguezbelmonte, A.M. Rodrigueztorres, E. Ramil, et al., Codon usage in *Kluyveromyces lactis* and in yeast cytochrome c-encoding genes, Gene 139 (1994) 43–49.

[13] S.E. Gharbia, J.C. Williams, D.M.A. Andrews, H.N. Shah, Genomic clusters and codon usage in relation to gene-expression in oral gram-negative anaerobes, Anaerobe 1 (1995) 239–262.

[14] P. Sharp, K. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, Nucleic Acids Res. 14 (1986) 5125–5143.

[15] C. Mathe, P. Rouze, Classification of *Arbidopsis thaliana* gene sequences: clustering of coding sequences into two groups according to codon usage improves gene prediction, J. Mol. Biol. 285 (1999) 1977–1991.

[16] L.I. Tong, C.T. Su, C.H. Wang, The optimization of multi-response problems in Taguchi method, Int. J. Qual. Reliab. Manage. 14 (4) (1997) 367–380.

[17] W.R. Dillon, M. Goldstein, Multivariate Analysis: Methods and Applications, Wiley, New York, 1984.

[18] SPSS, Advanced Statistical Analysis Using SPSS. Editor, Chicago, 2000.

[19] M.M. Tatsuoka, P.R. Lohnes, Multivariate Analysis, Macmillan Publishing Company, Inc., New York, 1998.

[20] M. Georgiopoulos, H. Fernlund, G. Bebis, G.L. Heileman, Order of search in Fuzzy ART and Fuzzy ARTMAP: effect of the choice parameter, Neural Networks 9 (5) (1996) 1541–1559.

[21] P.K. Simpson, Fuzz min–max neural networks—Part 2: clustering, IEEE Transactions on Fuzzy System 1 (1993) 32–45.

[22] B. Moore, ART 1 and pattern clustering, in: Proceedings of the 1988 Connectionist Models Summer School, 1989, pp. 174–185.

[23] G.A. Carpenter, S. Grossberg, D.B. Rosen, Fuzzy-ART: fast stable learning and categorization of analog patterns by an adaptive resonance system (original contribution), Neural Networks 4 (1991) 759–771.

[24] E. Helgason, O.A. Økstad, D.A. Caugant, H.A. Johansen, A. Fouet, M. Mock, I. Hegna, A.-B. Kolstø, *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence, Appl. Environ. Microbiol. 66 (2000) 2627–2630.

Kun-Lin Hsieh received Ph.D. degree in Industrial Engineering and Management, National Chiao Tung University, Taiwan. And, he received his BS in Computer Engineering from Chung-Yuan University, Taiwan and MS in Industrial Engineering from Yuan-Ze University, Taiwan. Presently, he is an assistant professor in the Department of Information Management, National Taitung University, Taiwan, ROC. His current research activities include IT and AI applications, quality engineering, process improvement, integrated analysis for DNA sequence.

I Ching Yang received his B.S. degree in 1996, M.S. degree in 1999, and Ph.D. degree in 2001, all from the National Cheng Kung University. In 2001, he joined the Department of Natural Education, National Taitung University. His current research interests include biological physics, statistical mechanics and black hole.