Grouping DNA sequences with similarity measure and application

Sanghyuk Lee
Dept. of Electrical and Electronic Engineering,
Xian Jiaotong-Liverpool University

Abstract Grouping problem with similarities between DNA sequences are studied. The similaritymeasure and the distance measure showed the complementary characteristics. Distance measure can be obtained by complementing similarity measure, and vice versa. Similarity measure is derived and proved. Usefulness of the proposed similarity measure is applied to grouping problem of 25 cockroach DNA sequences. By calculation of DNA similarity, 25 cockroaches are clustered by four groups, and the results are compared with the previous neighbor-joining method.

• Key Words : DNA sequence, similarity measure, fuzzy entropy

1. Introduction

DNA sequence analysis is an important work to analyze the logic of gene evolution. In order to analyze how far or similar they are among DNA sequences, similarity measure is proposed to calculate the degree of similarity measure. Hence, we consider the measure of similarity as the computing the distance between the species. It is well known that the DNA sequences only consist of four nucleotide bases {a, c, g, t}. However, there are numerous DNA bases, from 12-megabase yeast genome to 3-gigabase human genome. The inexact string matching algorithms of Needleman and Wunsch[1] and Smith and Waterman[2] have proven particularly useful for the quantifying the level of similarity between two sequences. In this literature, first we introduce the relation of similarity and distance measure, and propose the similarity and distance measure for computing distance from out-group DNA sequences. Similarity between two sets can be applied to the pattern classification or reliability field etc. Similarity measure has been known as the complementary meaning of the distance measure, i.e, $s+d=1$, where d and s are distance and similarity measure respectively. In the above, 1 means the sum of similarity and dissimilarity. In the previous literatures, fuzzy entropy of a fuzzy set represents a measure of fuzziness of the fuzzy set[3-10]. Furthermore, well-defined distance measure represents the fuzzy entropy. By the summing relation, we can notice that the similarity measure can be constructed through distance measure or fuzzy entropy function. Well known-Hamming distance is usually used to construct fuzzy entropy,so we compose the fuzzy entropy function through Hamming distance measure. Using the relation of distance measure and similarity measure, we construct the similarity measure with fuzzy entropy, and similarity measure is also constructed through distance measure. In the next section, the axiomatic definitions of entropy, distance measure and similarity measure of fuzzy sets are

*교신저자: Lee, S. (Xi'an Jiaotong-Liverpool University)
e-mail:Sanghyuk.Lee@xjtlu.edu.cn

접수일 2013년 07월 09일 수정일 2013년 07월 18일 게재확정일 2013년 07월 19일

introduced and fuzzy entropy is constructed through distance measure. In Section 3, similarity measures are constructed and proved through fuzzy entropy and the distance measure. Used distance measure is proposed by considering support average. To check the usefulness of the similarity measure, simple example is shown in Section 4. Conclusions are followed in Section 5. Notations of this paper are used with those of Liu's [6].

2. Preliminary

In this section, we introduce and discuss some preliminary results. Liu suggested three axiomatic definitions of fuzzy entropy, distance measure and similarity measure as follows [6]. By these definitions, we can propose entropy, and compare it with the result of Liu.

2.1 Some definitions of fuzzy entropy

In this subsection, we introduce some preliminary results about fuzzy entropy, distance measure, similarity measure, and related properties.

Definition 2.1 (Liu, 1992) A real function : $e:F(X)\to R^+$ or $e:P(X)\to R^+$ is called an entropy on $F(X)$, or $P(X)$ if e has the following properties:

(E1) $e(D)=0, \forall D\in P(X)$

(E2) $e([1/2])=\max_{A\in F(X)} e(A)$

(E3) $e(A^*)\le e(A)$, for any sharpening A^* of A

(E4) . $e(A)=e(A^C)$

where $[1/2]$ is the fuzzy set in which the value of the membership function is $1/2$.

Definition 2.3 (Liu, 1992) A real function $s:F^2\to R^+$ or $P^2\to R^+$ is called a similarity measure, if s has the following properties:

(S1) $s(A,B)=s(B,A),\ \forall A,B\in F(X)$

(S2) $s(A,A^C)=0,\ \forall A\in F(X)$

(S3) $s(D,D^C)=\max_{A,B\in F} s(A,B),\ \forall A,B\in P(X)$

(S4) $\forall A,B,B\in F(X)$, if $A\subset B\subset C$, then $s(A,B)\ge s(A,C)$ and $s(B,C)\ge s(A,C)$.

Liu also pointed out that there is an one-to-one relation between all distance measures and all similarity measures, that is $d+s=1$. Fuzzy normal similarity measure on F is also obtained by the division of $\max_{C,D\in F} s(C,D)$. If We divide universal set X into two parts D and D^C in $P(X)$, then the fuzziness of fuzzy set A be the sum of the fuzziness of $A\cap D$ and $A\cap D^C$. By this idea, following definition is followed.

From definition 2.1 and 2, we focus interesting area of universal set and extend the theory of entropy, distance measure and similarity measure of fuzzy sets. Fan and Xie derived new entropy via defined entropy, which is introduces by $e'=e/(2-e)$, where e is an entropy on $F(X)$.

2.2 Fuzzy entropy with distance measure

In this section, we propose entropy that is induced by the distance measure. Among distance measures, Hamming distance is commonly used -distance measure between fuzzy sets A and B,

$$d(A,B)=\frac{1}{n}\sum_{i=1}^{n}|\mu_A(x_i)-\mu_B(x_i)|$$

where $X=\{x_1,x_2,\cdots,x_n\}$, $|k|$ is the absolute value of k. Next Proposition shows that the distance relation of between fuzzy set and crisp sets.

Now we propose another fuzzy entropy induced by distance measure which is different from Theorem 3.1 of Fan, Ma and Xie [9]. Proposed entropy needs only A_{near} crisp set, and it has the advantage in computation of entropy.

Theorem 2.1 Let d be a σ-distance measure on $F(X)$ if d satisfies

$d(A^C, B^C) = d(A, B),\ A, B \in F(X)$, then

$$e(A) = 2d((A \cap A_{near}), [1]) + 2d((A \cup A_{near}), [0-2]) \quad (1)$$

is a fuzzy entropy.

Proofs of (1) are satisfied if (1) satisfy the Definition 2.1, so it is illustrated in [10]. Theorem 2.1 uses only A_{near} crisp set, hence we can consider another entropy. Which considers only A_{far}, and it has more compact form than Theorem 2.2.

Theorem 2.2 Let d be a σ-distance measure on $F(X)$ if d satisfies

$d(A^C, B^C) = d(A, B),\ A, B \in F(X)$, then

$$e(A) = 2d((A \cap A_{far}), [0]) + 2d((A \cup A_{far}), [1]) \quad (2)$$

is a fuzzy entropy.

In a similar way we can prove from (E1) to (E4) of Definition 2.1, it is also found in [10].

Proposed entropies Theorem 2.1 and 2.2 have some advantages to the Liu's, they use only one crisp sets A_{near} and A_{far}, respectively. Later we check the proposed entropy of Theorem 2.1 and 2.2 are the σ-entropy on $F(X)$ for any $A \in F(X)$, satisfying

$e(A) = e(A \cap D) + e(A \cap D^C)$.

3. Derivation of Similarity Measure

We obtain the fuzzy entropy with the distance measure in previous section. Generally, fuzzy entropy is expressed through distance measure, i.e., $e(A) = e(d(A))$. In our result, entropy is represented distance measure itself, $e(A) = d(A)$. Hence, by the result of Liu's,

$$d(A) + s(A) = 1 \quad (3)$$

we modify the similarity measure as $s(A) = 1 - e(A)$, that means fuzzy set*A*matches to the crisp set A_{near} nearly as $s(A)$ approaches to0. We illustrate the similarity measure with the entropy function in subsection 3.1 and the similarity measure construction using the distance measure in the subsection 3.2.

We propose the similarity measure in the following theorems. Theorem 3.1 is obtained by considering Theorem 3.2.

Theorem 3.1 For fuzzy set $A \in F(X)$, if d satisfies distance measure, then

$$s(A, A_{near}) = 4 - 2d((A \cap A_{near}), [1]) - 2d((A \cup A_{near}), [0]) \quad (4)$$

is the similarity measure between fuzzy set A and crisp set A_{near}.

Proofs are shown in reference 11. Similarly, we propose another similarity measure in the following theorem.

Theorem 3.2 For fuzzy set $A \in F(X)$ and distance measure d,

$$s(A, A_{near}) = 2 - 2d((A \cap A_{near}^C),[0]) - 2((A \cup A_{near}^C),[1]) \tag{5}$$

is the similarity measure of fuzzy set A and crisp set A_{near}.

Proofs are also shown in the reference 11. We have proposed the similarity measure that are induced from fuzzy entropy or distance measure.

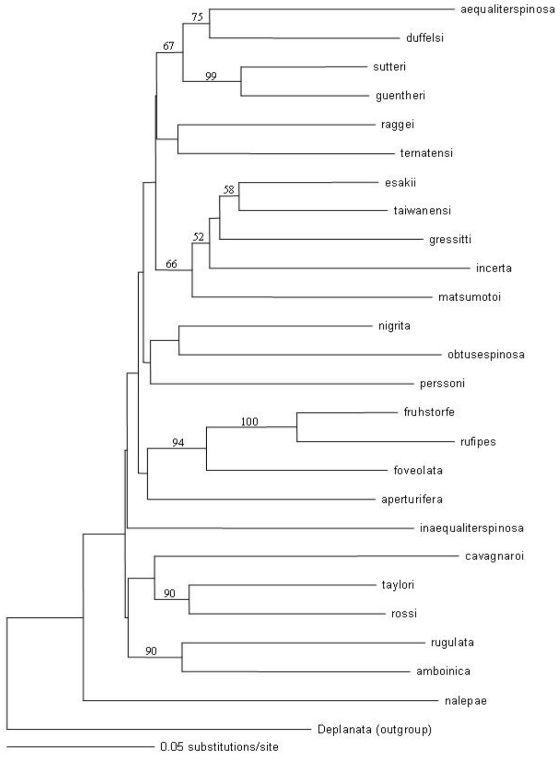
[Table 1] Species information used for the dataset.

Groups and species	Abbreviation of species name	Accession no.(Assigned no.)
S. raggei Roth	raggei	AB036206(18)
S. perssoni Roth	perssoni	AB036208(17)
S. aperturifera Roth	aperturifera	AB036209(3)
S. duffelsi	duffelsi	AB036210(5)
S. aequaliterspinosa	aequaliterspinosa	AB036216(1)
S. guentheri	guentheri	AB036220(10)
S. sutteri	sutteri	AB036221(22)
S. foveolata	foveolata	AB036222(7)
S. rufipes	rufipes	AB036223(20)
S. fruhstorferi	fruhstorferi	AB036224(8)
S. ternatensis	ternatensis	AB036226(25)
S. amboinica	amboinica	AB036228((2)
S. nigrita	nigrita	AB036230(15)
S. rugulata	rugulata	AB036231(21)
S. incerta	incerta	AB036232(12)
S. gressitti	gressitti	AB007529(9)
S. taiwanensis	taiwanensis	AB007527(23)
S. esakii	esakii	AB007518(6)
S. inaequaliterspinosa	inaequaliterspinosa	AB036234(11)
S. obtusespinosa	obtusespinosa	AB036236(16)
S. taylori	taylori	AB036239(24)
S. rossi	rossi	AB036240(19)
S. cavagnaroi	cavagnaroi	AB036241(4)
S. nalepae	nalepae	AB036242(14)
S. matsumotoi	matsumotoi	AB188688(13)
Miopanesthiadeplanata (outgroup)	deplanata	AB036104

4. Illustrative Example

The subsocial wood-feeding cockroach genus SalganeaStål (Blaberidae: Panesthiinae), including about 50 species, is distributed in the Indo-Malayan region and New Guinea of the Australian region. Since the completed COII gene sequences from about 25 species of the genus were already reported in reference[12], COII gene of the genus would be a good candidate to investigate patterns of sequence evolution and modeling within the lineage. Our dataset was constructed with the published COII (cytochrome oxydasesubsunit II) gene sequences [12]. According to a previous study, Miopanesthia Saussure is the basal group in the Panesthiinae. Thus the COII sequences of Miopanesthiadeplanata was used as out-group. The accession numbers and species names were summarized in Table 1 (also refer to Makewa et al., 2001). According to classical morphological studies (refer to Maekawa et al, 2001), the species of the genus for this study were classified as 4 groups of 18 species, but the other 7 species have not been unclassified yet. Firstly, we classified the dataset using the Neighbor-Joining Method and we applied our develop method for classifying the dataset. For the neighbor-joining analysis, we aligned the 685 sequences of the COII gene by using the Clustal X software. The gene sequences are aligned from 25 species of the genus Salganea and out-group. The 228 amino acids corresponding to the gene sequences were also used.

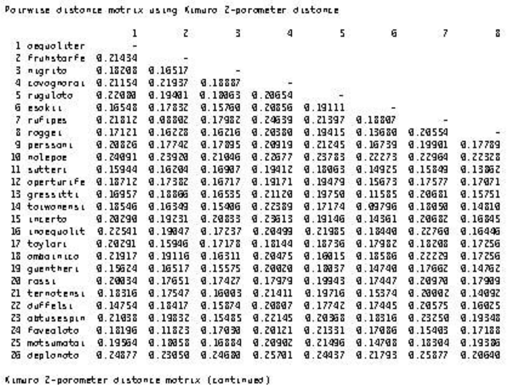
First, we carry out the analysis of the dataset by Neighbor-Joining Method. The phylogram tree induced by the neighbor-joining method is shown in Fig. 1. The number above and below the branches correspond to the percentage of 1000 bootstrap replicates. All nodes with no numbers are supported by 50% or less of the bootstrap values. Pairwise genetic distance based on Kimura 2-parameter is given to Table 2.



[Fig. 1] phylogram tree induced by the neighbor-joining method

With the similarity between DNA sequences, we try the unsupervised classification, then we exclude out-group.

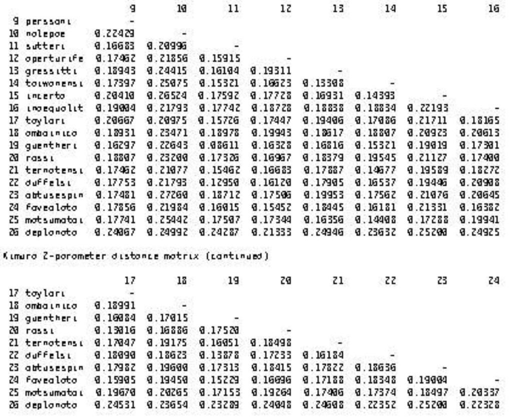
[Table 2] Genetic distance based on Kimura 2-parameter



Pairwise distance matrix using Kimura 2-parameter distance

		1	2	3	4	5	6	7	8
1	aequaliter	-							
2	fruhstorfe	0.21434	-						
3	nigrita	0.18208	0.16517	-					
4	cavagnaroi	0.21154	0.21937	0.18887	-				
5	rugulata	0.22080	0.19401	0.18063	0.20654	-			
6	esakii	0.16548	0.17832	0.15760	0.20856	0.19111	-		
7	rufipes	0.21812	0.08802	0.17982	0.24639	0.21397	0.18807	-	
8	raggei	0.17121	0.16228	0.16216	0.20380	0.19415	0.13680	0.20554	-
9	perssoni	0.20826	0.17742	0.17895	0.20919	0.21245	0.16739	0.19901	0.17789
10	nalepae	0.24091	0.23920	0.21046	0.22677	0.23783	0.22273	0.22964	0.22328
11	sutteri	0.15944	0.16204	0.16907	0.19412	0.18063	0.14925	0.15849	0.13862
12	aperturife	0.18712	0.17382	0.16717	0.19171	0.19479	0.15673	0.17577	0.17071
13	gressitti	0.16957	0.18806	0.16535	0.21120	0.19750	0.11585	0.20681	0.15751
14	taiwanensi	0.18546	0.16349	0.15406	0.22389	0.17174	0.09796	0.18050	0.14810
15	incerta	0.20290	0.19231	0.20833	0.23613	0.19146	0.14361	0.20682	0.16845
16	inaequalit	0.22541	0.19047	0.17237	0.20499	0.21985	0.18440	0.22760	0.16446
17	taylori	0.20291	0.15946	0.17178	0.18144	0.18736	0.17982	0.18208	0.17256
18	amboinica	0.21917	0.19116	0.16311	0.20475	0.16015	0.18586	0.22229	0.17256
19	guentheri	0.15624	0.16517	0.15575	0.20020	0.18037	0.14740	0.17662	0.14762
20	rossi	0.20034	0.17651	0.17427	0.17979	0.19943	0.17447	0.20970	0.17909
21	ternatensi	0.18316	0.17547	0.16003	0.21411	0.19716	0.15374	0.20002	0.14092
22	duffelsi	0.14754	0.18417	0.15874	0.20807	0.17742	0.17445	0.20575	0.16025
23	obtusespin	0.21038	0.19832	0.15485	0.22145	0.20368	0.18316	0.23250	0.19348
24	foveolata	0.18196	0.11823	0.17030	0.20121	0.21331	0.17086	0.15403	0.17188
25	matsumotoi	0.19564	0.10858	0.16884	0.20902	0.21496	0.14708	0.18304	0.19306
26	deplanata	0.24877	0.23050	0.24600	0.25701	0.24437	0.21793	0.25877	0.20640

Kimura 2-parameter distance matrix (continued)



		9	10	11	12	13	14	15	16
9	perssoni	-							
10	nalepae	0.22429	-						
11	sutteri	0.16683	0.20996	-					
12	aperturife	0.17462	0.21856	0.15915	-				
13	gressitti	0.18943	0.24415	0.16104	0.19311	-			
14	taiwanensi	0.17397	0.25075	0.15321	0.16623	0.13308	-		
15	incerta	0.20410	0.26524	0.17592	0.17728	0.16931	0.14393	-	
16	inaequalit	0.19004	0.21793	0.17742	0.18728	0.18838	0.18834	0.22193	-
17	taylori	0.20667	0.20975	0.15726	0.17447	0.19406	0.17086	0.21711	0.18165
18	amboinica	0.18931	0.23471	0.18978	0.19943	0.18617	0.18807	0.20923	0.20613
19	guentheri	0.16297	0.22643	0.08611	0.16328	0.16816	0.15321	0.19019	0.17301
20	rossi	0.18807	0.23200	0.17326	0.16967	0.18379	0.19545	0.21127	0.17400
21	ternatensi	0.17462	0.21077	0.15462	0.16683	0.17887	0.14677	0.19589	0.18272
22	duffelsi	0.17753	0.21793	0.12950	0.16120	0.17905	0.16537	0.19446	0.20908
23	obtusespin	0.17481	0.27260	0.18712	0.17506	0.19953	0.17562	0.21076	0.20645
24	foveolata	0.17856	0.21984	0.16015	0.15452	0.18445	0.16181	0.21331	0.16382
25	matsumotoi	0.17741	0.25442	0.17507	0.17344	0.16356	0.14408	0.17288	0.19941
26	deplanata	0.24067	0.24992	0.24287	0.21333	0.24946	0.23632	0.25200	0.24925

Kimura 2-parameter distance matrix (continued)

		17	18	19	20	21	22	23	24
17	taylori	-							
18	amboinica	0.18991	-						
19	guentheri	0.16084	0.17015	-					
20	rossi	0.13016	0.16886	0.17520	-				
21	ternatensi	0.17047	0.19175	0.16051	0.18498	-			
22	duffelsi	0.18090	0.18623	0.13878	0.17233	0.16184	-		
23	obtusespin	0.17982	0.19600	0.17313	0.18415	0.17822	0.18636	-	
24	foveolata	0.15905	0.19450	0.15229	0.16096	0.17188	0.18348	0.19004	-
25	matsumotoi	0.19670	0.20265	0.17153	0.19264	0.17406	0.17374	0.18497	0.20337
26	deplanata	0.24531	0.23654	0.23289	0.24048	0.24608	0.22352	0.25200	0.22328

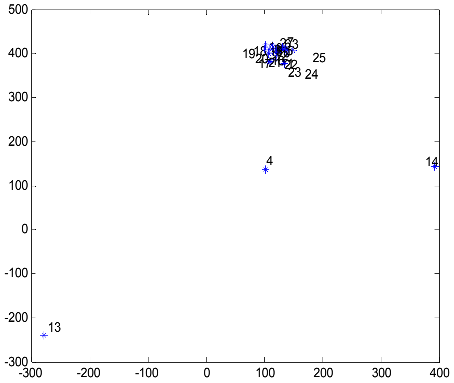
First we assign the 25 species to the successive numbers. Next we compute the distance from out group deplanata and fuliginosa to the 25 species as follows.

atgtcaacatgagctaatataggtacacaa … (deplanata)
atgacaacatgagccaacataaacttacaa … (aequaliterspinosa)

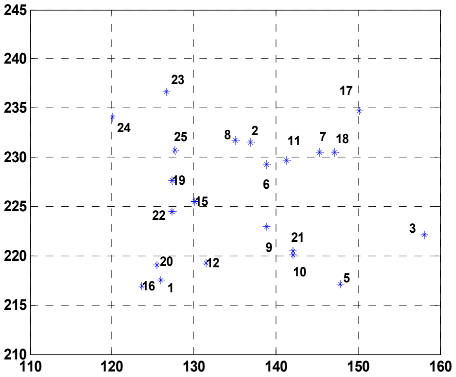
Distance from two outgroup can be defined as follows

$$D(\alpha,\beta,\gamma)=\sum_{i=1}^{n}(\alpha m(x_i-x_i')+\beta m(y_i-y_i')+\gamma m(z_i-z_i'))$$

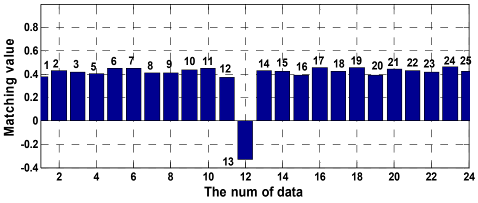
where, n denotes the number of amino acid, α, β, and γ are the weighting factors, and x_i, y_i, z_i be the successive amino acid out group. By the matching condition, $m(x_i-x_i')=1$, if $m(x_i-x_i')$. Otherwise satisfies -1. Hence 25 species can have the two distance value from two out-group, then25 species can be mapped into 2-dimensional plane with the assigned number in Table 1. Results are shown in Fig. 2. It shows that the distance from two out-groups. Besides of 4, 13, and 14, other 22 species are gatheredtogether, hence it is not easy to discriminate or classify. Magnification of 22 species point is illustrated in Fig. 3.



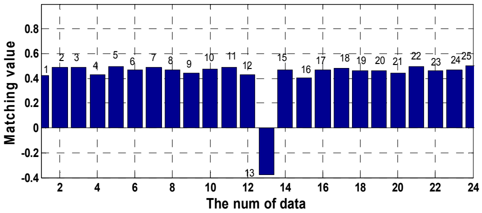
[Fig. 2] Distance from outgroup



[Fig. 3] Magnification of clustering area



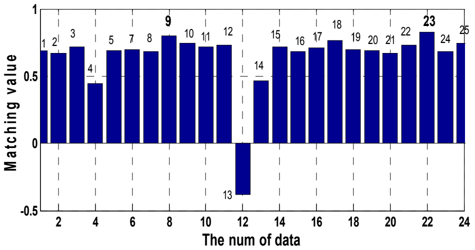
[Fig. 4] Similarity from cavagnaroi to other 24 species



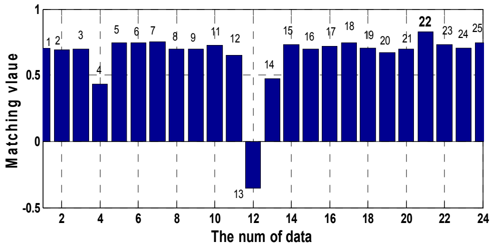
[Fig. 5] Similarity from nalepae to other 24 species

Autocorrelation values are normalized, and 25×¡¿25 data matrix is obtained. Next, we consider the correlation between DNA sequences. Low matching values are illustrated in Fig. 4 and 5. In these cases, alpha and beta are 1, and gamma denotes 0.2.

High matching values between DNA sequences are illustrated in Fig. 6 and 7.



[Fig. 6] Similarity from esakii to other 24 species



[Fig. 7] Similarity from guentheri to other 24 species

[Table 3] Species information used for the dataset

	Proposed Method
Group 1	1,5,10,22
Group 2	3,7,8,20
Group 3	6,9,12,15,16,18,23,25,
Group 4	2,4,11,13,14,17,,19,21,24

In this analysis, we consider that the multi matching condition. In any row, matching value over arbitrary threshing value can be chosen several. For example, 1 and 5 species has 0.739 maximum matching value in first row. 22 species has the largest matching value with 5 species in 5th row. 10 can be chosen in 22th row similarly. 22 species is also has the maximum value with 10 in 10th row. Hence, we can conclude that Group1, and Group2 are included in the same sectors [12].And elements of Group3 and Group4 are placed in near.

5. Conclusions

In order to classify data sets, evaluation of uncertainty and similarity was done by applying fuzzy entropy and similarity measure. Previous study on fuzzy entropy and similarity measure was introduced, and the derivation of similarity measure which can be represented by the function of distance measure. Proposed similarity measure and distance measure applied to the pattern recognition or data grouping. With the distance measure, 25 cockroach DNA sequences are clustered, and the results are compared with the previous one.

REFERENCES

[1] S.B. Needleman and C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Bio., Vol. 48, pp. 443-453, 1970.

[2] T.F. Smith and M.S. Waterman, "Identification of common molecular subswquences," J. Mol. Bio., Vol. 147, pp. 195-197, 1981.

[3] D. Bhandari, N. R. Pal, "Some new information measure of fuzzy sets," Inform. Sci. 67, 209-228, 1993.

[4] A. Ghosh, "Use of fuzziness measure in layered networks for object extraction: a generalization," Fuzzy Sets and Systems, 72, 331-348, 1995.

[5] B. Kosko, Neural Networks and Fuzzy Systems, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[6] Liu Xuecheng, "Entropy, distance measureand similarity measure of fuzzy sets and their relations," Fuzzy Sets and Systems, 52, 305-318, 1992.

[7] N.R. Pal, S.K. Pal, "Object-background segmentation using new definitions of entropy," IEEE Proc. 36, 284-295, 1989.

[8] J. L. Fan, W. X. Xie, "Distance measure and induced fuzzy entropy," Fuzzy Set and Systems, 104, 305-314, 1999.

[9] J. L. Fan, Y. L. Ma, and W. X. Xie, "On some properties of distance measures," Fuzzy Set and Systems, 117, 355-361, 2001.

[10] S.H. Lee, K.B. Kang and S.S. Kim, "Measure of fuzziness with fuzzy entropy function", Journal of Fuzzy Logic and Intelligent Systems, Vol. 14, No. 5, 642-647, 2004.

[11] S.H. Lee, J.M. Kim, and Y.K. Choi, "Similarity measure construction using fuzzy entropy and distance measure", LNAI Vol. 4114, 952-958, 2006.

[12] K. Maekawa, M. Kon, K. Araya and T. Matsumoto, "Phylogeny and Biogeography of Wood-Feeding Cockroaches, Genus SalganeaStål (Blaberidae: Panesthiinae), in Southeast Asia Based on Mitochondrial DNA Sequences", Journal Molecular Evolution, Vol. 53, pp. 651-659, 2001.

저자소개

Sanghyuk Lee

· 1988. Feb. : Chungbuk National University, Korea, Electrical Engineering, (B. Eng.)

· 1991. Feb: Seoul National University, Korea, Electrical Engineering, (M.S.)

· 1998. Feb: Seoul National University, Korea, Electrical Engineering, (Ph.D.)

· 2011. Aug.~Present: Professor of Dept. of Electrical Engineering, XI'anJiaotong-Liverpool University, China

· E-Mail : Sangyuk.Lee@xjtlu.edu.cn