Classification of species by information entropy and visualization by self-organizing map

Kentaro NISHIMUTA[1)], Ikuo YOSHIHARA[1)], Kunihito YAMAMORI[1)], Moritoshi YASUNAGA[2)]

[1)]1-1, Gakuen Kibanadai-nishi, Miyazaki, 889-2192, Japan, University of Miyazaki
(Tel : +81-985-58-7384; Fax : +81-985-58-7384)
(Email address : { nmuta, yoshiha, yamamori }@taurus.cs.miyazaki-u.ac.jp)
[2)]1-1-1, Tennoudai, Tsukuba, Ibaraki, 305-8573, Jaban
Graduate School of Systems and Information, Engineering,
University of Tsukuba

***Abstract*:** Most analysis methods of base sequences are based on the idea of pattern matching, but feature patterns known until today are supposedly a part of all the patterns hidden in base sequences. We develop a novel analyzing method based on disorder of base sequences, for example variance of data, 1/f fluctuation, information entropy and self-information. We try to classify species based on self-information of base sequences. Relations between species are visualized by self-organizing map (SOM). It is probable that neighboring regions of the SOM corresponds to near species. We compare neighboring regions of the SOM with neighboring branches of an evolutionary tree produced by the Clustal-W system.

Keywords: DNA, self-information, information entropy, self-organizing map

I. INTRODUCTION

Evolution degeneration of species must be remained in base sequences, because mutations have been accumulated in them for a long time. Various analysis methods on base sequences are investigated for revealing evolution of living creatures.

Most analysis methods of base sequences are based on pattern matching, but feature patterns known until today are probably a part of iceberg i.e. enormous kinds of patterns are ought to exist. Rule based analysis cannot be applicable to genes of unknown species.

We make up our mind to change the viewpoint of research from strict pattern matching to rough comparison. We propose a novel analyzing method using disorder of base sequences. Though disorder is somewhat vague, various features of species are reflected in disorder.

There exist many candidate methods of analysis concerning to disorder, for example, data compression ratio[1], long range order used in statistical mechanics[2], 1/f fluctuation[3], variance of data[4][5], information entropy[6] etc .

This article employs information entropy, especially self-information as an index of disorder. Self-information is calculated from appearance rate of codons in sub-sequences cut out from whole the base sequence.

We try to classify species based on self-information of base sequences without using pattern matching, and visualize relations of species by self-organizing map (SOM). One region represents one species.

We validate the method of classification of species and to compare neighboring regions on the SOM with neighboring branches of evolutionary tree produced by the Clustal-W system[7].

II. DISORDER ANALYISIS BY INFORMATION ENTROPY AND SOM

1. Composition of base sequences

Base sequences consist of four kinds of bases; A(Adenine), C (Cytosine), G (Guanine), and T (Thymine).

A sequence of three adjacent bases is called codon. The number of codons is $64=4^3$, because codon is a triplet of 4 kinds of bases.

2. Information entropy

Information entropy of species is assumed to be a measure of characteristic of species.

The complete event system *(X, P)* is determined by the occurrence probability P(X*i*), where X*i* is one of 64 codons. $\{X_i\}$ corresponds to AAA, AAG, AAC, … ,TTT.

$$\begin{pmatrix} X \\ P \end{pmatrix} = \begin{pmatrix} \mathrm{AAA} & \mathrm{AAG} & \cdots & \mathrm{TTT} \\ P(\mathrm{AAA}) & P(AAG) & , & P(TTT) \end{pmatrix} .$$

Self-information is defined by the probability P(X*i*), Self-information has the same number as the codon.

$$I(X_i) = -\log_2 P(X_i)\ (i = 1,2,\ldots,64)$$

Information entropy *H(X)* is an expected value of self-information *I(X)*, i.e. $H(X) = -P(X)\log_2 P(X)$.

$H(X)$ derives as follows,

$$H(X_1, X_2, \cdots, X_n) = -\sum_{i=1}^{64} P(X_i)\log_2 P(X_i).$$

3. Self-organizing map (SOM)

Self-organizing map (SOM) proposed by Kohonen is an unsupervised-leaning neural network which is used for clustering and visualization [8].

The codon of self-information as vector of 64th dimension is used for input of SOM. SOM has two-layer structure. Input layer has nodes the number of which is the same as the number of dimension of the input.

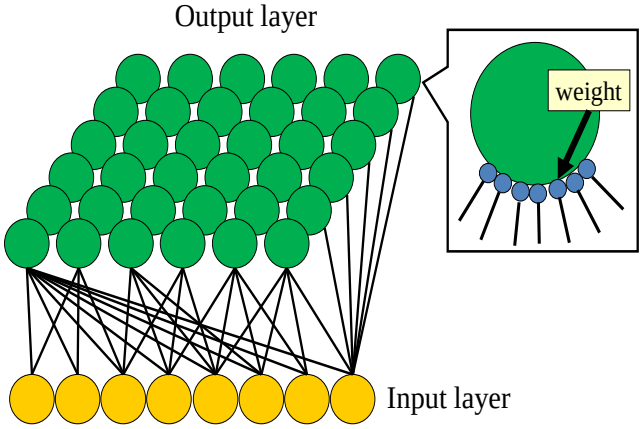


Fig 1 Illustration of input and output layers of SOM

The output layer of SOM is allocated on rectangle lattice(x,y). The size of output layer is arbitrarily decided.

Weights connect input layer and output layer. Weight is strength of connection between input layer and output layer. Two layers unite completely by weight (Fig 1). The relation between the output*(O)* ,weight*(W)* and the input*(I)* is shown by the following expressions.

$$O(x, y) = \sum_{i=1}^{64} W_i I_i .$$

We use SOM to distinguish species by self-information.

4. Positional information entropy

Sub-sequences are parts of a base sequence cut out with a given length of bases and are information entropy a positional function of the base sequence (Fig 2).

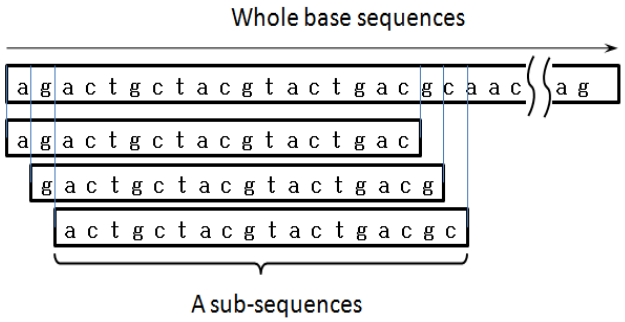


Fig 2 Base representation of the whole base sequence

Self-information of codons is calculated by the appearance probability of codons in sub-sequence (Fig 3).

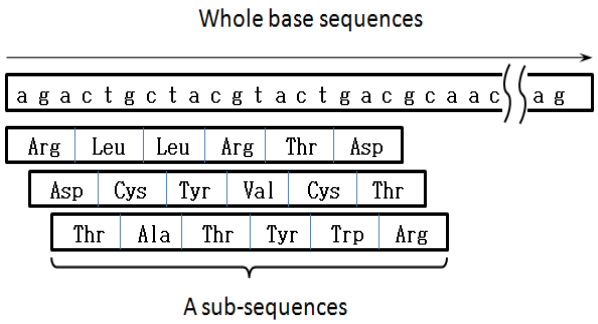


Fig 3 Codon representation of sub-sequences

III. EXPERIMENTS

1. Experimental condition

We validate the proposed method by the following steps.

[Step 1]. Calculation of self-information.

[Step 2]. Visualization by SOM.

[Step 3]. Comparison SOM with evolutionary tree.

Experimental data of base sequences are ribosomal protein genes of 20 species in Table 1, which are taken from the Ribosomal Protein Gene (RPG) Database established by Frontier Science Research Center, University of Miyazaki[9].

The ribosomal proteins used are as follows;

RPL3 RPL4 RPL5 RPL6 RPL7 RPL7A
RPL8 RPL9 RPL10 RPL10A RPL17 RPLP23

2. Pre-experiment

First of all, we need to determine experimentally an appropriate length of sub-sequence. We conduct small scale experiments will give us nearly optimal parameters of SOM. 6 kinds of species are used; { A(Hs), B(Mm), H(Ag), I(Ce), J(Mg), and O(Um) } in Table 1.

We change the length of sub-sequence from 64 bases to 2048 bases. Sub-sequence cannot cut out more than 2048 bases, because length of base sequence is different because of species. Parameters of experiments are as follows; map-size is 25×25, and iteration of learning is 100.

3. Results of pre-experiment

Length of sub-sequence 2048 gives good figure of SOM(Fig 4). Species are separated in 6 regions. Length of sub-sequence is long, species of 6 kinds all separated.

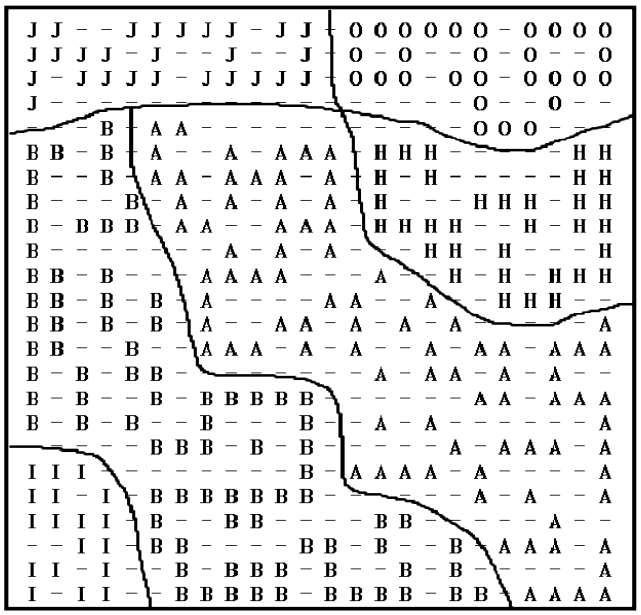


Fig 4 Result of pre-experiment (6 species)

4. Experiment

We validate the proposed method employing data of 20 species. Length of sub-sequences is same for all species i.e.2048 bases. Parameters in experiments are as follows; map-size is 40×40, and iteration of learning is 100.

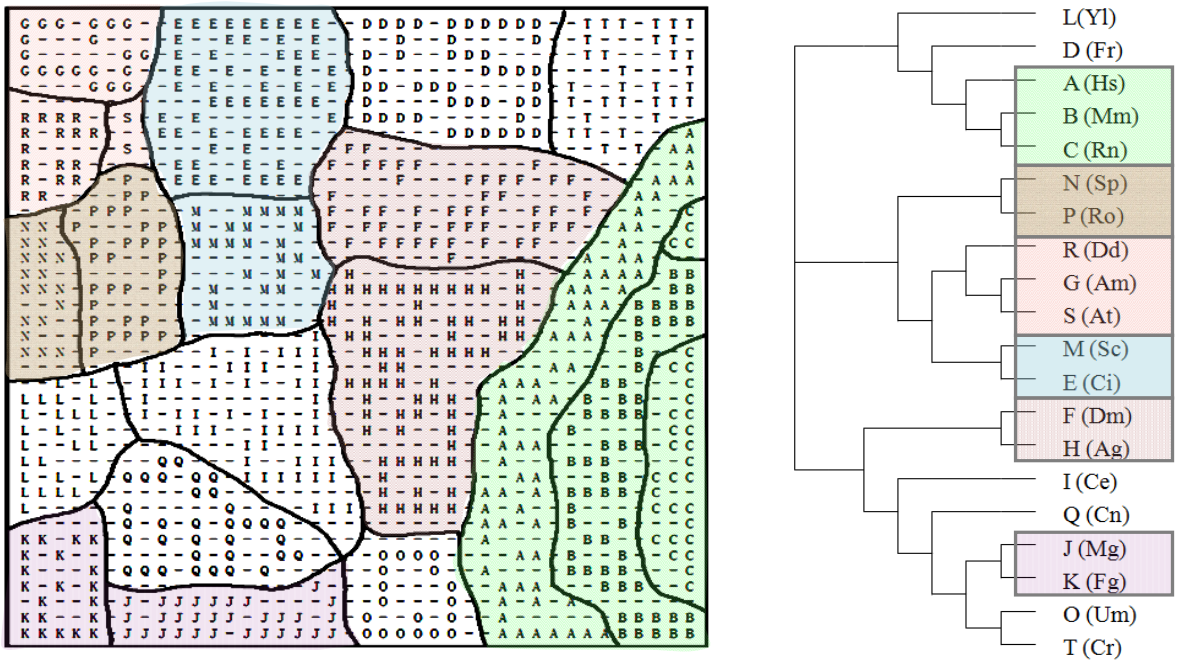
5. Results and discussions

Fig 5 (A) is a map produced by SOM and Fig 5 (B) is an evolutionary tree produced by the Clustal-W system. We take a look of parts with same colored mask pattern.

The map of SOM is separated in 21 regions. There are 2 regions of C(Rn). We confirmed to correspond between the neighboring regions of the SOM to the neighboring branches of the evolutionary tree.

Table 1 Correspondence table for experiment

symbol	abbr	species	symbol	abbr	species	symbol	abbr	species	symbol	abbr	species
A	Hs	H.sapiens	F	Dm	D.melanogaster	K	Fg	F.graminearum	P	Ro	R.oryzae
B	Mm	M.musculus	G	Am	A mellifera	L	Yl	Y.lipolytica	Q	Cn	C.neoformans
C	Rn	R.norvegicus	H	Ag	A.gamiae	M	Sc	S.cerevisiae	R	Dd	D.discoideum
D	Fr	F.rubripes	I	Ce	C.elegans	N	Sp	S.pombe	S	At	A.thaliana
E	Ci	C.intestinalis	J	Mg	M.grisea	O	Um	U.maydis	T	Cr	C.reinhardtii



(A) (B)

Fig 5 Comparison of SOM with evolutionary tree

IV. CONCLUSION

We proposed an analyzing method of base sequences employing disorder. The method is emposed of calculation of self information from base sequences of species and visualization by SOM. A rate of SOM is to separate all the species into different regions. We compare neighboring regions of SOM with neighboring branches of evolutionary tree produced by the Clustal-W system.

The neighboring regions of the SOM are corresponding to the neighboring branches on evolutionary tree. The experiments lead us to the conclusion that the analysis with disorder and SOM is useful for classifying species.

REFERENCES

[1] I. Yoshihara, H. Takashima, K. Yamamori, K. Sugawara (2006),"Similarity Comparisons of Seeds Using Image Compression Technology", Memoirs of the Faculty of Engineering, University of Miyazaki,35,251-256

[2] Y. Sakaguchi, I. Yoshihara, K. Yamamori, N. Kenmochi (2006),"Criterion of Evolution Analyzing DNA Sequences based on Statistical Mechanics" Memoirs of the Faculty of Engineering, University of Miyazaki,35,269-274

[3] Y. Koyama, K. Nishimuta, K. Yamamori, I. Yoshihara, M. Yasunaga (2010), "Quest for genetic information hidden behind disorder in DNA sequences", Proc. of 15th Int. Symp. on Artificial Life and Robotics (AROB'15),824-827

[4] K. Yamamori, Y. Fujita, M. Aikawa, I. Yoshihara (2007), "Identification of Exon-intron Boundaries by Integration of Base-oriented Genetic Programming and Statistical Heuristics", Proc. of 12th Int. Symp. on Artificial Life and Robotics (AROB'12), 657-660

[5] T. Abe, S. Kanaya, M. Kinouchi (2003) "Informatics for Unveiling Hidden Genome Signatures", Genome research, 13, 693-702

[6] S. Miyazaki, H. Sugawara, M. Ohya (1996), "The efficiency of entropy evolution rate for construction of phylogenetic trees", Genes & Genetic Syst, vol.71,5,323-327

[7] Clustal-W system; http://clustalw.ddbj.nig.ac.jp/

[8] T. Kohonen (2005), Self-Organizing Maps. Springer-Verlag Tokyo. (Translated into Japanese by Tokutaka et.al.)

[9] RPG database; http://ribosome.med.miyazaki-u.ac.jp/