# Quest for Genetic Information Hidden behind Disorder in DNA Sequences

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

*[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 : { koyama, yamamori, yoshiha }@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 of conventional base sequences are analyzed by order of base sequences, for example, pattern matching. Pattern matching compares unknown base sequences with that of known gene to find similar patterns and to identify gene information. We try to search for hidden information in DNA sequences without pattern matching. We focus on disorder of base sequences, because disorder analysis is available, if we do not know particular function of genes. We use the exponent α of $1/f^{\alpha}$ fluctuation and self-information as indices of disorder. Our experimental data are ribosomal protein of eukaryotic species. The exponent α is calculated for three kinds of data, i.e. whole base sequences, base sequences in exon or intron. The average of α in exon regions are smaller than that in intron regions. It suggests that exon regions are somewhat more ordered than intron regions. SOM is used to look for similarity of species by self-information which is calculated for codons of base sequences. SOM shows that self-information is usable for a classification of species.

***Keywords***: DNA, 1/f, fluctuation, self-information, entropy, SOM

## I. INTRODUCTION

DNA of organisms involves various genetic information to form themselves, which has been inherited from generation to generation.

Recent years, it has been active to analyze DNA sequences for finding organic evolution. Pattern matching is widely used to find similar patterns in a pair of base sequences and to identify function and evolutionary relationships.

We notice using disorder of base sequences instead of order to reveal hidden genetic information. Disorder analysis is available, even if we do not know meaningful patterns in advance. We focus on $1/f^{\alpha}$ fluctuation of chaos theory and self-information to extract hidden information buried in disorder of base sequences.

Takushi and Miyagi [1-3] researched $1/f^{\alpha}$ fluctuation of bacteriophage φ-X174 and suggested that several fractal group correlations exist in a sub-sequences and whole sequences. They examine whether specific expression of genes can be found. We aim at finding hidden information in DNA sequences of many species.

Information theory is introduced by Shannon can play a major role in analyzing features and characteristics of sequences, if DNA sequences carry the information over generations. Since information entropy and self-information are measure of the uncertainty associated with a random variable. They are treated as an amount of characteristic of DNA sequence.

## II. DISORDER ANALYISIS BY $1/f^{\alpha}$ FLUCTUATION

### 1. $1/f^{\alpha}$ Fluctuation

" $1/f$ fluctuation", strictly speaking " $1/f^{\alpha}$ fluctuation" is observed in natural phenomena such as flame of candle, the breeze and the classical music etc. We try to use $\alpha$ as an index of disorder. It is said in musical sphere that $\alpha$ of classical music is 0.5-1.75 and that of rock music is 0.01-1.0[4]. $\alpha$ is probably depending on kinds of music. We use analogy to explain disorder of base sequences by $\alpha$ .

## 2. Calculation of exponent $\alpha$

The exponent $\alpha$ of $1/f^{\alpha}$ fluctuation of base sequences is calculated by the following procedure.

**Step1.** Numerical expression of DNA sequences

DNA sequences consist of four kinds of bases; A (Adenine), C (Cytosine), G (Guanine), and T (Thymine). It is necessary to convert all bases into numerical values before Fourier transform. Bases are converted into complex numbers, which are located almost on four apexes of a square.

$$\begin{aligned}
&\mathrm{A}\cdots(1+r_1)+(1+r_2)i\\
&\mathrm{G}\cdots(-1+r_3)+(1+r_4)i\\
&\mathrm{C}\cdots(1+r_5)+(-1+r_6)i\\
&\mathrm{T}\cdots(-1+r_7)+(-1+r_8)i \qquad r_n\in(-0.01,0.01)
\end{aligned}$$

**Step2.** Discrete Fourier Transform (DFT)

DFT is employed for Fourier Transform of discrete data. Power spectrum $S_f$ is calculated by Eq.(2), where $f_n(n=0,1,\cdots,N-1)$ are base sequences expressed by complex number mentioned above.

$$F_k=\sum_{n=0}^{N-1} f_n e^{-i\frac{2\pi kn}{N}}\ (k=0,1,\cdots,N-1). \qquad (1)$$

$$S_f=|F_k|^2. \qquad (2)$$

**Step3.** Calculation of $\alpha$

$\alpha$ is a gradient in low-frequency range of the power spectra (Fig.1). The gradient is calculated by linear regression.

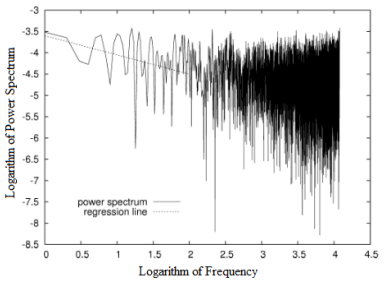


Fig.1. Power spectrum and regression line

## 3. Experiments

*A. Experimental Conditions*

We use ribosomal protein data of eukaryotic species for experiments. Table.1 shows the names of the eukaryotic species and their abbreviation (abbr.). The kinds of the ribosomal proteins are as follows;

RPS2 RPS5 RPS7 RPS8 RPS9 RPS12
RPL5 RPL8 RPL9 RPL26 RPL32 RPLP0

These data have been released on Ribosomal Protein Gene Database: (RPG) by Frontier Science Research Center, University of Miyazaki [5]. We calculated $\alpha$ for exon regions, intron regions and whole regions including exon and intron for all the data. Base sequences involving no less than 50 bases are used.

Table.1. Eukaryotic species

| name of species | abbr. |
|---|---|
| H.sapiens | Hs |
| M.musculus | Mm |
| A.ganbiae | Ag |
| C.elegans | Ce |
| P.falciparum | Pf |
| M.grisea | Mg |

*B. Results*

Fig.2-a, Fig.2-b and Fig.2-c show the average $\alpha$ of exon, intron and whole regions. Fig.2-a and Fig.2-b show that average of $\alpha$ in exon regions are smaller than that in intron regions in all species. In the field of music, $\alpha$ of rock music are smaller than that of classical music. We assume rock musics correspond to exon and classical musics correspond to intron. Rock music is stormy sounds whereas classical music is comparatively calm sounds. Exon regions may involve partly distinguished sub-sequences. Exons are made for protein–coding transcripts, but introns are not. The experiments suggest that exon regions probably have more particular base sequences than intron regions.

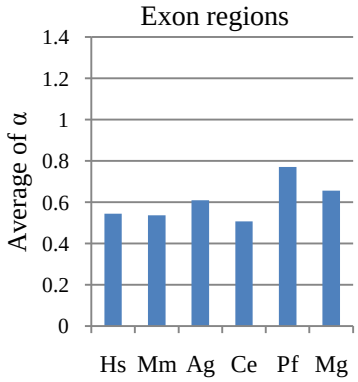


Fig.2-a. Average $\alpha$ for exon regions

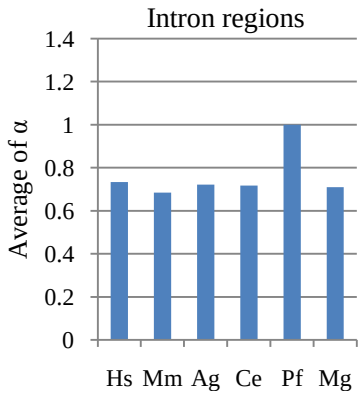


Fig.2-b. Average $\alpha$ for intron regions

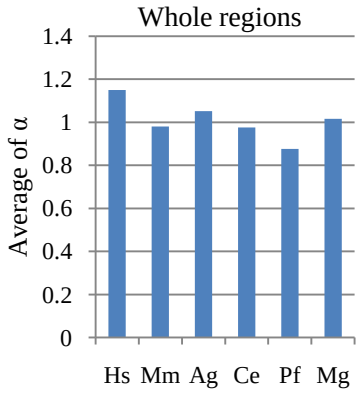


Fig.2-c. Average $\alpha$ for whole regions

## III. DISORDER ANALYISIS BY INFORMATION ENTROPY AND SOM

### 1. Information entropy

Information entropy of base sequences is assumed to be a measure of characteristic of species. Information entropy is defined as follows. DNA sequences $X$ is composed of four kinds of bases {A, G, C, T}. The complete event system $(X, P)$ of $X$ is determined by the occurrence probability $P(X)$ for each event $X(\mathrm{A}), X(\mathrm{G}), X(\mathrm{C}), X(\mathrm{T})$.

$$\begin{pmatrix} X \\ P \end{pmatrix} = \begin{pmatrix} \mathrm{A} & \mathrm{G} & \mathrm{C} & \mathrm{T} \\ P(\mathrm{A}) & P(\mathrm{G}) & P(\mathrm{C}) & P(\mathrm{T}) \end{pmatrix}.$$

Self-information is defined by the probability $P(X)$,

$$I(P(X)) = -\log_2 P(X). \qquad (3)$$

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

Because $\{X_i\}$ corresponds to A, G, C, T, $H(P(X_i))$ becomes as follows,

$$H(P(x_1), P(x_2), \cdots, P(x_n)) = -\sum_{i=0}^{n} P(x_i)\log_2 P(x_i). \qquad (4)$$

The number of occurrence is $64=4^3$, because codon is triplet of 4 kinds of bases. When we analyze the base sequences from the view point of codon.

Fig.3. is information entropy of codon in whole regions. The figure does not show clear difference in six species. An additional approach is required.

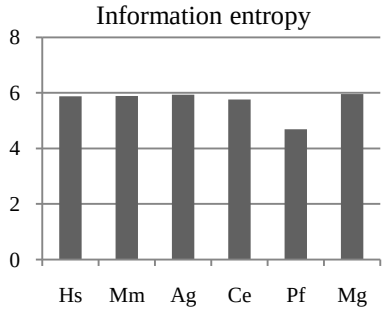


Fig.3. Information entropy of codon

### 2. Positional self-information

*Sub-sequence* is a part of base sequences cut out with a given number of bases and depends on the position of the base sequences (Fig.4). An additional approach is to calculate self-information for codons of sub-sequences, that varies from the beginning of the base sequence to the end (Fig.5).

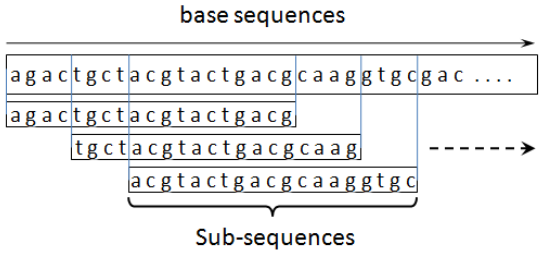


Fig.4. Shifting of cut-out position

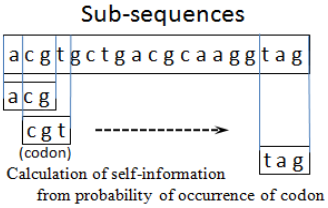


Fig.5. Shifting of codon

### 3. Self-organizing map (SOM)

Self-organizing map (SOM) proposed by Kohonen [6] is an unsupervised-leaning neural network which is used for clustering and visualization. SOM is also an algorithm to map high dimensional data to two dimensional space, to visualize buried structure in high dimensional space on two understandable dimensional maps.

We use SOM to distinguish species by self-information. The number of elements of input vectors to SOM is 64, corresponding to variety of codons.

### 4. Experiments

*A. Experimental Conditions*

Experimental data are the same as used by $1/f^{\alpha}$ fluctuation. Parameters of SOM in experiments are as follows; map-size is $30 \times 30$, and iteration of learning is 100.

Table.3 shows the number of training samples and correspondence symbols of species on the map in Fig.6. Lengths of sub-sequences are same for all samples i.e. 2048 bases.

Table.3. Correspondence list for SOM

| name of species | #s of training samples (RPL32) | symbol |
|---|---|---|
| H.sapiens | 234 | H |
| M.musculus | 188 | M |
| A.ganbiae | 119 | A |
| C.elegans | 68 | C |
| P.falciparum | 65 | P |
| M.grisea | 81 | m |
| Total | 755 | |

*B. Results and discussion*

Fig.6 shows all the species are classified. We change the length of the sub-sequence from 64 bases to 2048 bases. The experiments show the longer, the better. In case of 2048 bases, all the species are successfully clarified shown in Fig.6.

## VII. CONCLUSION

We try to search for hidden information in DNA sequences with disorder of DNA sequences, because this method is available, even if characteristic pattern is unknown. We use ribosomal proteins of eukaryotic species and investigate the exponent $\alpha$ of $1/f^{\alpha}$ fluctuation and self-information as indices of disorder.

The exponent $\alpha$ is calculated for exon regions, intron regions and whole regions. The average of $\alpha$ in exon regions are smaller than that in intron regions. It suggests that exon regions are somewhat more ordered than intron regions.

SOM successfully distinguishes species with self-information for codons of sub-sequences in the whole region.

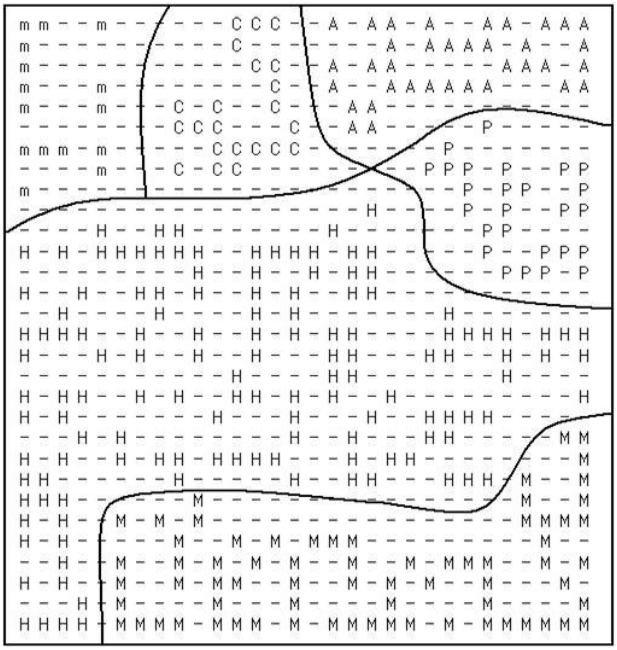

Fig.6. SOM of RPL32

## REFERENCES

[1] Takushi E and Miyagi H (2000), Fractal Packing of the DNA Sequence of Bacteriophage φ-X174 (in Japanese). Bull. Fac. Sci. Univ. Ryukyus, 70:43-46
[2] Takushi E and Miyagi H (2001), Fractal Packing of the DNA Sequence of Bacteriophage φ-X174 (II). Bull. Fac. Sci. Univ. Ryukyus, 71:21-23
[3] Takushi E and Miyagi H (2001), Fractal Packing of the DNA Sequence of Bacteriophage φ-X174 (III). Bull. Fac. Sci. Univ. Ryukyus, 72:43-47
[4] 1/f of music; http://bodysonic.cc/1fyuragi.htm
[5] RPG; http://ribosome.med.miyazaki-u.ac.jp/
[6] Kohonen T (2005), Self-Organizing Maps (Translated in to Japanese by Tokutaka et.al.). Springer-Verlag Tokyo.