1/f ゆらぎによる生物進化の解析

佐藤 真和[1)] 吉原 郁夫[2)] 山森 一人[3)] 安永 守利[4)]

Analysis of Biological Evolution using 1/f Noise

Masakazu SATO Ikuo YOSHIHARA Kunihito YAMAMORI
Moritoshi YASUNAGA

ABSTRACT

The eukaryotic genome consists of two kinds of regions. One is coding region that operates to synthesize protein. The other is non-coding region that does not operate to synthesize protein. Most of DNA base sequences of eukaryotic genome are non-coding region. To find embedded information in non-coding region is one of the most important tasks for clarifying evolution of species.

We analyze DNA base sequences of *Dictyostelium Discoideum* and ribosomal protein gene with 1/f noise in chaos theory.

Key Words:
1/f noise, Dictyostelium Discoideum, ribosomal protein gene

1 はじめに

近年、DNA 一次構造におけるヌクレオチドの塩基配列を決定することが出来るようになり、いろいろな生物種の遺伝子について塩基配列データが急速に集積されつつある。しかし、それらの情報がどのような意味をもっているのかという問題については十分な成果があげられていない。そのため、遺伝情報の意味を抽出する有効な方法が求められている[9)]。

真核生物のゲノムは大きく分けて二つの領域から構成されている。一つは、タンパク質合成に直接寄与するコード領域であり、その他は、タンパク質合成に直接は寄与しない非コード領域である。真核生物の塩基配列のほとんどは非コード領域である。非コード領域の塩基配列に隠れていると思われる進化の痕跡を見つけることは、生物の進化を明らかにするための重要な課題である。しかし、この領域の機能については、まだまだ不明な点が多い。

従来、Li and Kaneko[4)]、Voss[5)]、Stanley[6)] などにより DNA 塩基配列に $1/f^{\alpha}$ ゆらぎが存在することが確認された。彼らの結論によると塩基配列のパワースペクトルに長距離相関が存在し、低周波領域が $1/f^{\alpha}$ で近似できる。この指数 α はパワースペクトルを低周波側で近似する直線の傾きを表しており、長距離相関の度合いを示す。

また、それをもとに澤岻、宮城ら[1)2)3)]は、バクテリオファージ ϕ-$X174$ 遺伝子で塩基数の累積を取りながらパワースペクトルを求めると、$1/f^{\alpha}$ ゆらぎの指数 α に一次転移のような変化が確認され、がんや突然変異、イントロンの発現性の可能性などを指摘した。さらに彼らは、p51 がん抑制遺伝子やバクテリオファージ ϕ-$X174$ 遺伝子の部分配列の傾き α と全塩基配列の傾き α にはフラクタル的な相関が存在することを示した。彼らは、基底 2 の FFT (Fast Fourier Transform) を用いているため部分配列の長さは 64、128、256、512 などの値を用いている。本研究では、様々な長さの部分配列長を扱うために、基底（2,3,5）の混合基底 FFT プログラムを用いる。混合基底 FFT を用いることで、標本長が 2^n のような粗い間隔に制約されることなく、より綿密な解析が可能となる。

本研究では、$1/f^{\alpha}$ ゆらぎの指数 α を配列の乱雑さの尺度として用いて、不規則性の観点から特徴・類似度の調査を行う。本研究では、代表的なモデル生物である細胞性粘菌の非転写領域データ、また１０種の生

[1)]工学研究科情報工学専攻学生
[2)]情報システム工学科教授
[3)]情報システム工学科助教授
[4)]筑波大学電子・情報工学系教授

物のリボソームタンパク質遺伝子の転写領域データを使用する。

2 DNAとタンパク質

DNAは、ヌクレオチドとよばれる分子が結合してできた高分子である。ヌクレオチドはそれぞれ1個の糖、1個のリン酸、1個の塩基からなっている。ヌクレオチドには4種類のものがあるが、塩基の部分だけが異なっており、それぞれ、アデニン（A)、グアニン（G)、シトシン（C)、チミン（T)と呼ばれている。DNAは糖とリン酸でできた主鎖に沿って特定の配列で塩基が並んだものである。

DNAの翻訳領域の情報に従ってタンパク質が生成されることを遺伝子の発現という。転写開始点より上流は非転写領域とよばれ、プロモーターやエンハンサーなどの遺伝子発現を制御する情報などが記されている。転写の最初の段階ではDNAを鋳型として、相補的なRNAが作られる。DNAとRNAは分子の種類としてはほとんど同じであって、塩基対の相補性も完全に成り立つ（ただし、4種類の塩基のうち、DNAのチミンがRNAではウラシルに変わっている)。できたRNAには一般にはエクソンとイントロンという二つの部分があり、交互に並んでいる。そして、次のスプライシング過程でイントロン部分だけが切り離され、エクソンだけがつながったメッセンジャーRNAができる。その翻訳領域の情報に従ってアミノ酸の種類が決定され、タンパク質が合成される[9]。

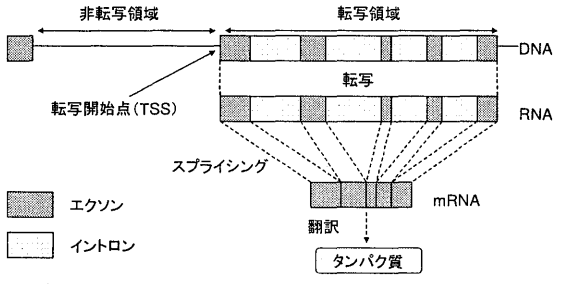

図. 1 タンパク質の生成過程

3 1/fゆらぎ

3.1 $1/f$ ゆらぎとは

ゆらぎとは、量や質の変化が一定の周期を示しているように見えるが、わずかなずれを生じる予測できない変化の事である。このゆらぎの中で、パワースペクトルが周波数に反比例し、傾きが－1となるゆらぎは $1/f$ ゆらぎと呼ばれる。この $1/f$ ゆらぎは、ろうそくの炎、そよ風、小川のせせらぎなどの様々な自然現象の中や生体リズムで確認された。一般的に、様々な傾きがあるため $1/f^{r}$ と表す。本研究では、$1/f^{r}$ の指数 α を用いてDNA塩基配列の解析を行う。

傾きが0 ($\alpha = 0$) の場合、周波数とスペクトルが無関係すなわちホワイトノイズである。傾きがきつくなればなるほど、相関性が高いといえる。傾きが -1 ($\alpha = 1$)、つまり $1/f$ ゆらぎは規則性と非規則性が適当にまじりあっていると考えられる。

3.2 $1/f^{\alpha}$ の指数 α の計算方法

本研究では、$1/f^{\alpha}$ の指数 α を用いて塩基配列の解析を行う。指数 α は以下の手順で算出する。

Step1 塩基配列を数値に変換する（データの前処理）

Step2 パワースペクトルを算出する

Step3 最小周波数から23ポイントまでの低周波領域を、最小二乗法を用いて線形回帰する (図. 2)

Step4 求まった回帰直線の傾きを $-\alpha$ とする

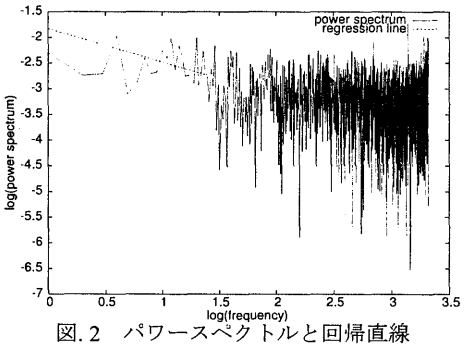

図. 2 パワースペクトルと回帰直線

3.2.1 塩基配列の数値化

DNA塩基配列はA（アデニン)、G（グアニン)、C（シトシン)、T（チニン）の4つの塩基から構成されている。フーリエ変換を実行する前に各塩基を数値に変換しなければならない。本研究では、図. 3のように各塩基が対称に配置されるように複素数を用いる。また、パワースペクトルが頻繁にゼロになるのを防ぐため、小さな値を足しておく。

- A$\cdots(1 + r_1) + (1 + r_2)i$
- G$\cdots(-1 + r_3) + (1 + r_4)i$
- C$\cdots(1 + r_5) + (-1 + r_6)i$
- T$\cdots(-1 + r_7) + (-1 + r_8)i$

$r_n \in (-0.1, 0.1)$

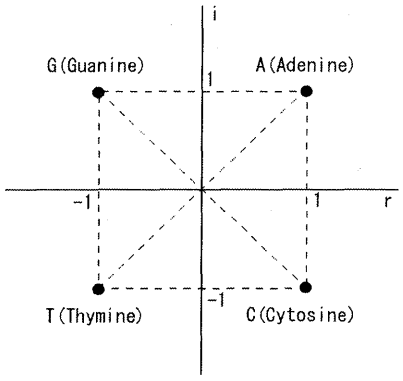

図. 3 塩基配列の数値化

3.2.2 フーリエ変換とパワースペクトル

標本長 N の離散データ $g_k(k = 0, 1, \cdots, N-1)$ の離散フーリエ変換 (Discrete Dourier Transform : DFT) は以下の式で定義される。

$$G(n) = \frac{1}{N}\sum_{k=0}^{N-1} g(k)W_N^{nk} \quad (n = 0, 1, \cdots, N-1) \qquad (1)$$

$$W_N = exp(-j * 2\pi/N)$$

N : 標本長
$g(k)$: k 番目のデータ
W_N : ひねり因子

次に、パワースペクトルは以下の式で定義される。

$$S(n) = |\, G(n)\,|^2 \qquad (2)$$

本研究では、計算時間短縮のために FFT を用いる。

4 非転写領域の実験

4.1 モデル生物 : 細胞性粘菌

細胞性粘菌は原始的な真核生物の一種であり、代表的なモデル生物である。その生活環（ライフ・サイクル）は、無性生殖環と有性生殖環からなる。本研究では、無性生殖環の DNA 塩基データを用いるので、ここでは無性生殖環だけについて説明する。細胞性粘菌には発現の時期により、大きく分けて 4 つの発現期がある。(図. 4)

(1) アメーバ状で独立して生活し、増殖する増殖期 (V stage)

(2) 集合体の中心に向かい集合する集合体期 (A stage)

(3) ナメクジ状の移動体期 (S stage)

(4) 子実体と呼ばれる構造体を形成する子実体期 (C stage)

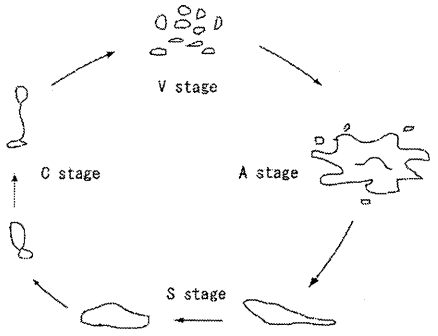

図. 4 細胞性粘菌のライフサイクル

細胞性粘菌はこのような特徴的な生活様式に加えて、培養が容易で外的条件を調整しやすいことなどの好条件を備えている。そのため、細胞分化・形態形成のメカニズム、細胞集合、細胞間通信などの生理・生化学、遺伝学、微細構造学の有力なモデル生物として用いられている [7)8)]。

4.2 実験データと実験方法

非転写領域の細胞性粘菌データを使用して実験を行う。これらのデータは、共同研究を行っている筑波大学から提供していただいた。718 個の細胞性粘菌データの中から、4 つの発現期のうち 1 つの発現期でよく発現するデータだけを抽出した (表. 1)。各データはデータ長 2100bases であり、非転写領域（2000 bases)、転写領域（100 bases）で構成されている。また、2000 番目が転写開始点 (Transcriptional start site : TSS) である。

表. 1 細胞性粘菌データ数

ステージ名	データ数
V stage	131
A stage	86
C stage	54
S stage	61

長さ L の部分配列を移動させながら、各部分配列の α を算出する (図. 6)。例えば、Nbases の塩基配列の場合、$1 \sim L$、$2 \sim L+1$、$\cdots$、$N-L+1 \sim N$ まで部分配列を移動させる。つまり、一つの塩基配列に対して $N-L+1$ 個の部分配列の α を求める。

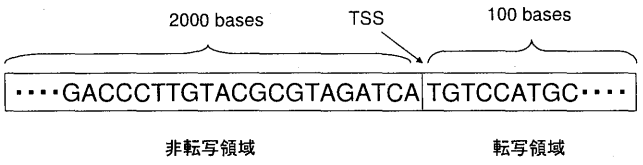

図. 5 細胞性粘菌データ

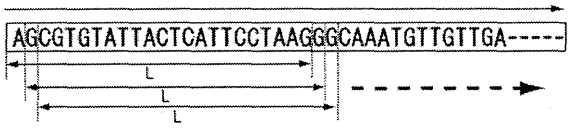

図. 6 実験方法

4.3 結果と考察

4.3.1 部分配列の $1/f^{\alpha}$ の指数 α の変化

図. 7 は、$L = 48$ のときの、各発現期毎の α の変化を比較している。X 軸は部分配列の起点を表し、$X = 2000$ が転写開始点である。グラフを見やすくするために、各ステージの結果を Y 軸方向にいくらかシフトしている。例えば、図. 7 では A ステージの $\alpha + 0.6$、C ステージの $\alpha + 0.4$、S ステージの $\alpha + 0.2$、Y 軸方向にシフトしている。V ステージ、C ステージでは、他のステージと比較して転写開始点前で α の平均が増加しているのが分かる。これは、転写開始点前の各塩基の出現頻度の変化が影響しているかもしれない。ただし、平均する前の各データで、このような増加がはっきりとみられるわけではない。また、TSS 直後の転写領域では、どのステージでも α が減少している。

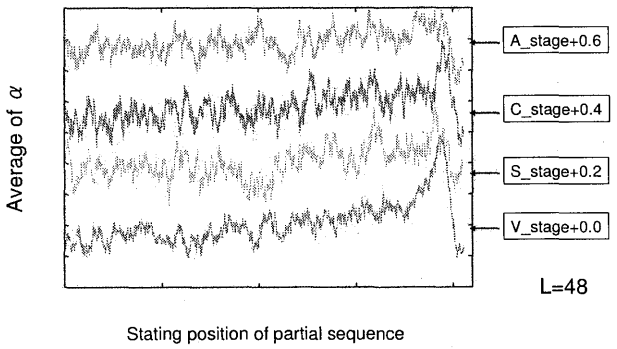

図. 7 各ステージの α の変化比較 (L=48)

次に、L を変化させたときの結果を、図. 8 に示す。ここで、結果の順序は図. 7 と対応している。L を変化させることで、それぞれ違った特徴がいくつか確認された。例えば、

- V stage : TSS 直前の増加
- A stage : $X = 700$ 付近の減少 ($L = 64$、81)
- C stage : いくつか特徴的な増減 ($L = 162$)
- S stage : TSS の前での二箇所や $X = 600$ 付近での一部増加 ($L = 45$)、$X = 600$ 付近 ($L = 162$) での減少

これらの結果より、配列長を変化させることによって、各ステージの特徴や違いをうまくとらえているように思える。

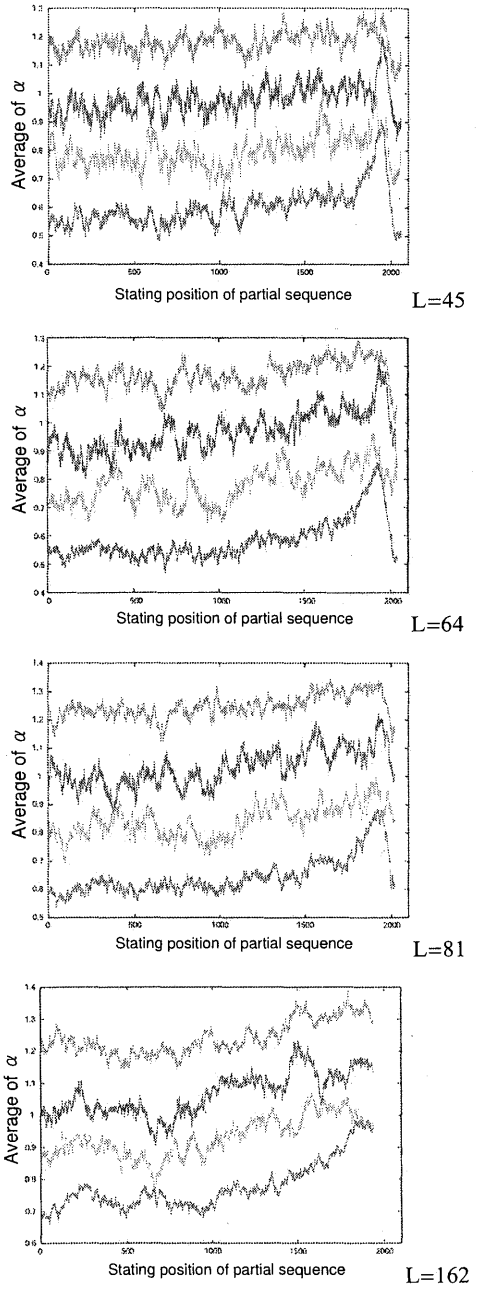

図. 8 各部分配列長 (L) での α の変化

5 転写領域の実験

5.1 リボソームタンパク質

リボソームはmRNAの翻訳を担う重要な細胞内装置である。ヒトを含む高等生物において、リボソームは4種類のRNAと79種類のタンパク質からできている。これらの各成分の発現は協調的に制御されており、その構造は酵母や細菌に至るまで大変よく保存されている。生物にとって安定したリボソームの供給は、細胞の維持や増殖、またさまざまな刺激に応じたタンパク質生成に必須となる。そのため、リボソームに生じた変異はただちに細胞に障害をもたらし、それが個体における発生の異常、ひいては死を招くと考えられる[10]。

リボソームは、生物種に共通して存在することや、エクソンの塩基配列の保存性が高いことから、生物進化、特にイントロンに内在する進化の痕跡を明らかにするための重要な研究材料として注目されている。

5.2 実験データ

転写領域のリボソームタンパク遺伝子データを使用して実験を行う。これらのデータは、宮崎大学フロンティア科学実験総合センターによってRibosomal Protein Gene database : RPG†で公開されている。実験では、表. 2に示す10種の生物のデータを使用する。なお、各データはエクソンとイントロンが交互に並んでいる(図. 9)。また、データ長はデータによって異なる。

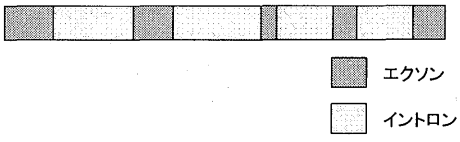

図. 9 リボソームタンパク質データ

表. 2 リボソームタンパク質データ数

略称	データ名	データ数
Hs	ヒト	80
Mm	マウス	79
Dm	ショウジョウバエ	99
Ce	センチュウ	88
Sc	出芽酵母	138
Sp	分裂酵母	141
Pf	マラリア原虫	86
Mj	メタン性古細菌	62
Ec	大腸菌	54
At	シロイヌナズナ	226

†URL http://ribosome.miyazaki-med.ac.jp/

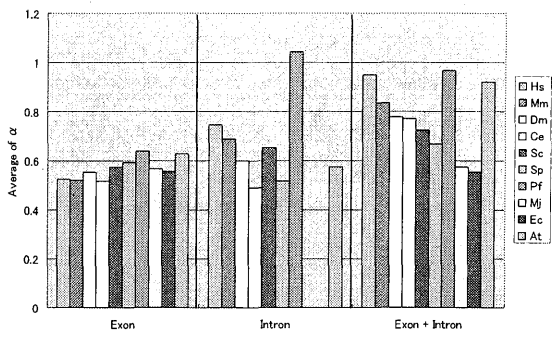

図. 10 配列全体の α

5.3 結果と考察

5.3.1 配列全体の比較

まず、配列全体の α の比較を行う。各データに対し α を計算する。また、エクソン、イントロン部分を切り出し比較を行う。ただし、データ長が23以下のデータは削除する。そして、生物種毎に平均をもとめる。ここで、標本長は各生物種のデータ長と等しい。

図. 10は、生物種毎のエクソン、イントロン、転写領域(エクソン+イントロン)での比較を表す。メタン性古細菌(Mj)と大腸菌(Ec)は原核生物であるため、イントロンが存在しない。

エクソンでの比較ではイントロンほど生物種の違いはあまりみられない。リボソームタンパク質遺伝子は、エクソン領域の塩基配列は保存性が高く生物種であまり違いがないためだと考えられる。逆にイントロンと転写領域では、おおよそ、高等生物に近い生物ほど α が高いようにみえる。マラリア原虫(Pf)で α が高いのは塩基配列中に繰り返しパターンが多く存在するため、規則的にみえるためだと考えられる。転写領域の大部分はイントロンであるので、転写領域の結果はイントロンの結果の影響を受けていると思われる。

5.3.2 部分配列の $1/f^{\alpha}$ の指数 α の変化

データ長($L = 25$〜279)の部分配列をシフトさせながら α がどのように変化していくか、また特徴はないか実験を行う。ここでは、生物種共通のオーソロガス遺伝子を比較する。オーソロガス遺伝子とは、異なる種において構造上、機能上類似な遺伝子である。図の X 軸は部分配列の起点、Y 軸は α である。そこで、興味深い特徴を見つけた。二つの生物種を比較する際に、X軸方向へいくつかシフトすると図. 11・図. 12のように類似した部分があることが分かった。リボソームタンパク質は、エクソンの保存性が非常に高いので、その影響であるとも考えることができる。しかし、こ

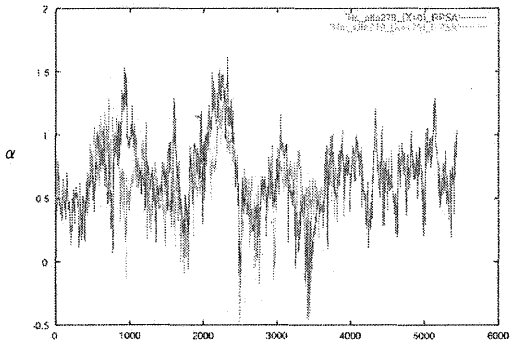

図. 11 Hs と Mm(X+479) の α の変化 (RPSA:L=279)

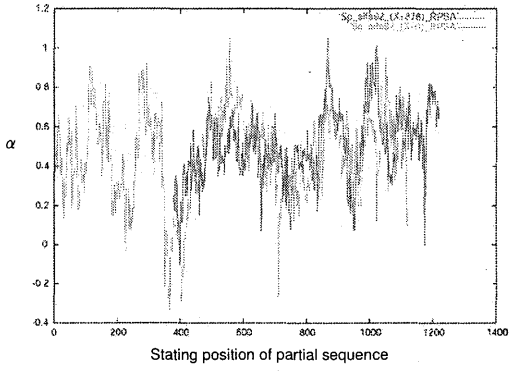

図. 12 Sp(X+376) と Sc の α の変化 (RPSA:L=92)

のことが生物進化の解明のための手掛かりとなる可能性もある。

またこれらの結果は、二つの生物種を比較する際に、もっとも適した部分配列長 (L)、かつ X 軸方向へいくつシフトすればよいかを相互相関を用いて検出した。

6 おわりに

$1/f^{\alpha}$ ゆらぎの指数 α を配列の乱雑さの尺度として用いて、不規則性の観点から特徴・類似度の調査を行った。本研究では、標本長が 2^n のような粗い間隔に制約されることなく、より綿密な解析を行うために混合基底 FFT プログラムを用いた。

まず、細胞性粘菌の非転写領域データを使用して、発現期ごとの特徴や違いを調査した。部分配列長を変化させることで、それぞれ各ステージごとに違った特徴がいくつか確認された。本手法により各ステージの違いや特徴をうまくとらえることができたと思われる。

次に、リボソームタンパク質の転写領域データを使用して、10種の生物種の塩基配列の比較を行った。そこでは、エクソンでは生物種によってあまり α の差がみられないがイントロンと転写領域では高等生物に近い生物ほど α が高いことがわかった。これは、リボソームタンパク質遺伝子では、異なる生物種でエクソンの保存性が高いという特徴と一致している。また、イントロンで生物種によってある程度系統樹にそった違いがあるという事は、イントロンに進化の痕跡が内在されていると考えられる。さらに、各生物のオーソロガス遺伝子を比較したところ、α の変化が類似している部分があるという興味深い結果を得た。

今後の課題として、コドン単位での塩基配列の数値化、部分的に類似した位置の自動検出方法の提案があげられる。

7 謝辞

本研究の一部は文科省科研費・基盤 (C)No.17500146 により行われた。

参考文献

[1] 澤岨英正, 宮城拓, "バクテリオファージ ϕ-X174 : DNA 塩基配列のパワースペクタルにおけるフラクタル的充填," *Bull.Fac.Sci.,Univ.Ryukyus* ,No. 70, pp.43-46, (2000).

[2] E. Takushi and H. Miyagi, "Fractal Packing of the DNA Sequence of Bacteriophage ϕ-X174 (II)," *Bull.Fac.Sci.,Univ.Ryukyus* ,No. 71, pp.21-23, (2001).

[3] E. Takushi and H. Miyagi, "Fractal Packing of the DNA Sequence of Bacteriophage ϕ-X174 (III)," *Bull.Fac.Sci.,Univ.Ryukyus* ,No. 72, pp.43-47, (2001).

[4] W. Li and K. Kaneko, "DNA correlations," *Nature* ,Vol.360, pp.635-636, (1992).

[5] R. F. Voss, "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences," *phys.Rev.Lett.*, , Vol.68 , pp.3805-3808, (1992).

[6] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," *Nature* , Vol.356, pp.168-170, (1992).

[7] 吉原郁夫, 山口崇, 山森一人, 安永守利, "1/f ゆらぎを用いた細胞性粘菌の特徴抽出," *Memoirs of the Fac. of Engineering, Miyazaki Univ.*, Vol.32, pp.277-282, (2003).

[8] T. Onitani, I. Yoshihara, K. Yamamori, M. Yasunaga, "Extraction of Feature Patterns Embedded in Non-Transcribed Region of Dictyostelium Discoideum," *SEAL '04*, SWP-8, No. 123, (2004).

[9] 美宅成樹, 金久實, "ヒトゲノム計画と知識情報処理," 培風館, 1995.

[10] 剣持直哉, "リボソームと疾患" 実験医学, Vol.22, No.17 (増刊), (2004).