$1/f$ ゆらぎを用いた細胞性粘菌の特徴抽出

吉原郁夫[1] ・ 山口崇[2] ・ 山森一人[3] ・ 安永守利[4]

Feature extraction of Cellular Slime Molds using $1/f$ fluctuation

Ikuo YOSHIHARA[1],Takashi YAMAGUCHI[2],Kazuhito YAMAMORI[3],Moritoshi YASUNAGA[4]

Abstract

DNA analysis of Cellular Slime Molds are important for investigating human genome. The exponent α of the $1/f^{\alpha}$ fluctuation is used as an index of irregularity to extract features from DNA sequence. It specially pays attention to the part before and after the transcription starting point in the DNA sequence of the Sellular Slime Molds. A difference in the transcriptional region and the untranscriptional region is found by this method. $1/f$ fluctuation reveals a repetition pattern in the DNA sequence.

Key Words:

$1/f$ fluctuation, power spectrum, FFT, dictyostelium discoideum, gene, cDNA, transcription start point, transcriptional region, untranscriptional region

1 はじめに

近年、ヒトの全DNA（デオキシリボ核酸）情報を解読するヒトゲノム解析計画が終了するなど、様々な遺伝子の解析が急速に進んでいる [1]。しかしながらその解析データ の活用法についての研究はまだまだ発展途上である。一方、生物システムのさまざまな挙動に1/f揺らぎと呼ばれる自然システムに広く観測されるゆらぎが観測されてきている。

細胞性粘菌は有性生殖をする生物の中で最も原始的な真核生物である。そのためヒト等の遺伝子に比べてDNAの量は少ないが、ヒトにも共通する特徴を持っている可能性が高い。

粘菌の無性的ライフサイクルには単細胞生物のように振舞う時期と多細胞生物ように振舞う時期がある。この特徴のためヒト等の高等で複雑な多細胞生物の発生メカニズムを解明するための単純なモデル生物として研究されている。

Peng[2],Li and Kaneko[3][4],Voss[5] 等はDNA塩基配列の長距離相関の存在を示した。それによると塩基配列を数値列に置換しフーリエ変換することによって求めたパワースペクトルの低周波数部分は、関数$1/f^{\alpha}$によって近似することができるという。このことから、長距離相関は$1/f^{\alpha}$ゆらぎとも呼ばれ、$1/f^{\alpha}$の指数αはパワースペクトルの低周波側を近似する直線の傾きを表しており、長距離相関の傾きαともいう。

また、それをもとに澤岻 [6][7][8] はバクテリオファージϕ-174のDNA塩基配列の部分配列に長距離相関$1/f^{\alpha}$ゆらぎが存在することを示した [6][7][8]。さらにそのうちの幾つかの部分配列と全DNA塩基配列との間にフラクタル的パッキングが暗号も含めてなされていることを指摘している [6][7][8]。

本研究は細胞性粘菌の転写開始点以降とその前の非転写開始点にどのような違いがあるのか、不規則性の観点から見て細胞性粘菌の各発現期毎に違いがあるか調べることを目的とする。細胞性粘菌のDNAデータの中でも研究のあまり進んでいない非転写領域、その中でも遺伝子の発現に関わると思われる転写開始点の近くについて不規則性の観点から違いをみる。また、本研究では遺伝子データ列の規則性の度合の指標として、$1/f^{\alpha}$の指数αを使用する。

2 遺伝のメカニズム

DNAは生物の遺伝情報を保持しており、細胞中では染色体の形で保存されている。DNA上にはタンパク質を生成するための情報が並んでいるが、全てがそういった情報なわけではない。DNA全体でタンパク質を生成するための情報は全体の3％程である。

2.1 DNAと遺伝子

遺伝子とはDNAの中でタンパク質を生成するための情報を持った部分である (図1)。遺伝子は保持しているタンパク質情報が必要になった時に読み取られる。そして遺伝子発現を経てタンパク質を生成

[1] 宮崎大学情報システム工学科教授
[2] 宮崎大学情報工学科学生
[3] 宮崎大学情報システム工学科助教授
[4] 筑波大学電子・情報系助教授

する。

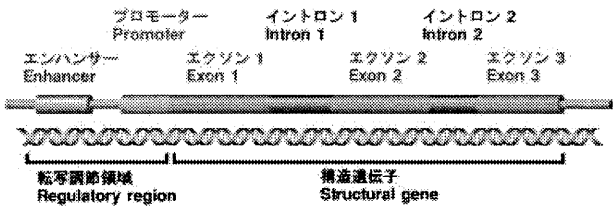

図 1. DNA 構造 (参考文献 [9] より)

2.2 遺伝子の選択

ある生物の DNA が各細胞とも共通であるのに、細胞により構造や機能は異なる。これは個々の細胞で必要な遺伝子だけが発現されるという機構のためである。

DNA の中には遺伝子でなく転写されない非転写領域の中に遺伝子調節領域というものがある (図 2)。ここに基本転写因子や遺伝子調節タンパク、RNA ポリメラーゼが結合し転写が始まる。この遺伝子調節領域によって発現する遺伝子が決定されているものと考えられている。しかし遺伝子調節領域をはじめとする非転写領域の機能については解っていないことが多い。

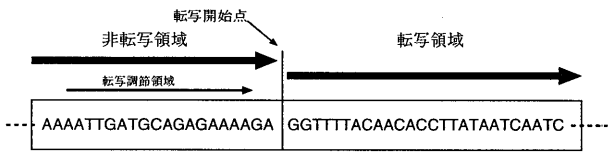

図 2. 転写領域と非転写領域

2.3 遺伝子の発現のメカニズム

まず細胞の核の中で DNA 上の遺伝子のコピーが生成される。このコピーは RNA と呼ばれる物質で DNA とよく似た一本鎖の核酸である。RNA には mRNA、tRNA、rRNA の 3 種類がある。

(1) mRNA: 事実上の、タンパク質を作るための遺伝子のコピー

(2) tRNA: タンパク質合成に必要な材料 (アミノ酸) を運ぶ RNA

(3) rRNA: タンパク質合成をするリボソームを構成する RNA

これらの RNA が作られる過程を転写という。

次いで mRNA が細胞核の外に出る。細胞核の外に出た mRNA はリボソームと結合し、コピーされた情報の通りにタンパク質が作られる。このタンパク質を生成する過程を、翻訳という。

2.4 cDNA

cDNA とは DNA の内、タンパク質合成に不要なイントロン部分を取り除いた mRNA を、人工的に写しとったものである (図 3)。

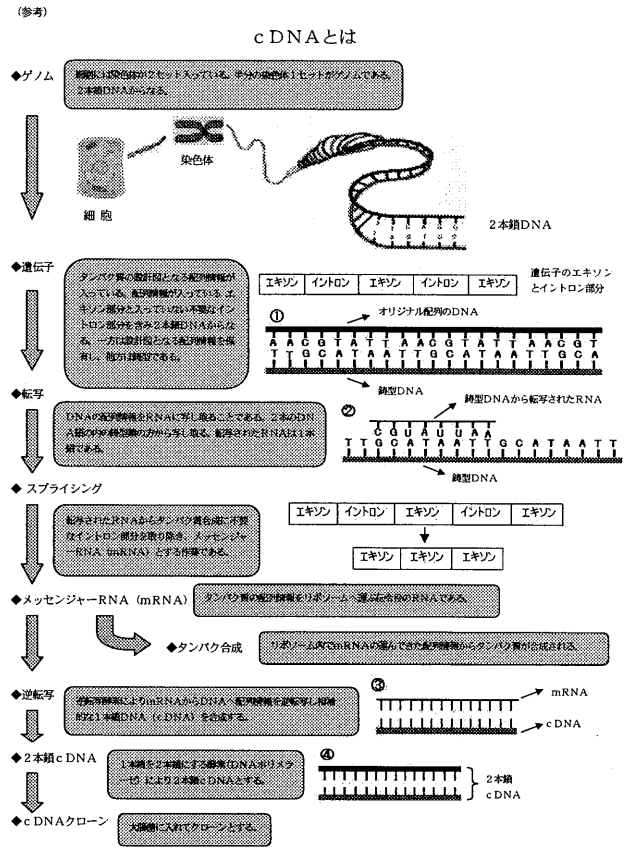

図 3. cDNA (参考文献 [10] より)

3 細胞性粘菌

近年、ゲノムの分野で細胞性粘菌の研究が盛んに行われているのは、真核生物の中でもヒトなどと比べ DNA 構造が単純なため、特徴を掴むのが容易だろうと考えられるからである。また細胞性粘菌はクラウン生物群よりも前に分岐している真核生物である。そのため、その後に分岐するクラウン生物群の生物にも同じ特徴がある可能性がある。

3.1 発現期

細胞性粘菌の無性的ライフサイクルには 4 つの発現期 (図 4) がある。

(1) vegetative stage (アメーバ状の増殖期)
細胞性粘菌はこの時期独立して生活、増殖する。増殖していきある程度の密度になり、栄

養物質を使い尽くすと aggregating stage に移行する。

(2) aggregating stage (粘液アメーバの集合体)
中心に向かい流れをなして集合する。集合体は全体が粘質物質に包まれていき、slug stage に移行する。

(3) slug stage (ナメクジ状の移動体)
移動しながら予定柄細胞と予定胞子細胞の細胞選別を行い、のちに停止し culminating stage に移行する。

(4) culminating stage (子実体)
slug stage で行った細胞分別で予定胞子細胞となったものが乳頭突起物となり、さらに残りの部分が基部となり最終的に子実体と呼ばれる多細胞体を構成する。

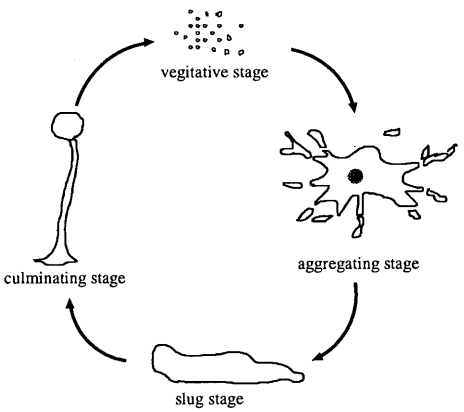

図 4. 細胞性粘菌の無性的ライフサイクル

4 $1/f^{\alpha}$ ゆらぎ

ゆらぎとは，巨視的には一定であっても、微視的には平均値前後で絶えず変動している現象である。

ゆらぎの中で，パワースペクトルを解析しグラフの傾きが－１となるゆらぎを１／ｆゆらぎと呼ぶ.

一般に、傾斜がきつくなればなるほど、不規則な変化が抑制されて大きな変化が支配するようになり、次を予測しやすい現象と考えられる。それに対して、傾斜が緩やかになればなるほど、不規則な変化が頻発するため、次を予測しにくい波形になる。

本研究では遺伝子データ列の規則性の度合の指標として $1/f^{\alpha}$ 指数 α を使用する。これはこれまでの「データにどのような並びがあるのか」という規則性をを探すのとは逆の不規則性の観点からデータを見つめる事により、分析したいとの動機からである。また、生命体のさまざまなところからゆらぎが観測されている [11] ことから遺伝子データからも観測できるのではないかと考えた。

4.1 パワースペクトル

信号の振り幅を一定の周波数帯域毎に分割し、各帯域毎に周波数の関数として表したものをパワースペクトルという。本研究ではフーリエ変換によって、時間軸波形から周波数軸波形を求める。

N 個の離散データ $x_n(n = 1, 2, \cdots, N)$ の離散フーリエ変換 $X(f)$ は

$$X(f) = \sum_{n=1}^{N} x_n e^{\frac{-j2\pi nf}{N}} dt \tag{1}$$

となる。

このときパワースペクトル $S(f)$ は

$$S(f) = |X(f)|^2 \tag{2}$$

とすることができる。

しかしながら計算の高速化のため本実験では離散フーリエ変換のかわりに FFT を使用する。

4.2 FFT

離散フーリエ変換を計算するには N^2 回の乗算が必要であり，N が大きくなると計算に必要な時間は急速に増大する。ところが,1965 年にクーリーとチューキーによって提案された高速フーリエ変換 (Fast Fourier Transform, 略して FFT) のアルゴリズムは, これを最小で $N \log_2 \frac{N}{2}$ 回にまで減らす。

式１は

$$X_f = \sum_{n=1}^{N} x_n W^{nf}, \quad W = e^{\frac{-j2\pi}{N}} \tag{3}$$

と書き換えるとができる。

N が偶数である場合を考える. $f = 2m$, 即ち X_f の偶数番目の成分に注目したとき, $W^N = 1$ である事を利用して,

$$X_{2m} = \sum_{n=1}^{\frac{N}{2}} x_n W^{2mn} + \sum_{n=\frac{N}{2}+1}^{N} x_n W^{2mn} \tag{4}$$

$$= \sum_{n=1}^{\frac{N}{2}} x_n W^{2mn} + \sum_{n=1}^{N} x_{n+\frac{N}{2}} W^{2mn+mN} \tag{5}$$

$$= \sum_{n=1}^{\frac{N}{2}} (x_n + X_{n+\frac{N}{2}}) W^{2mn} \tag{6}$$

と表せる。この時、

$$x_n^E \equiv x_n + x_{n+\frac{N}{2}} \tag{7}$$

とおくなら、式 6 は $X_{2m}(m = 1, \cdots, \frac{N}{2})$ が長さ $\frac{N}{2}$ の数列 x_n^E のフーリエ変換となる事を表わしている。

同様にして,$f = 2m+1$, 即ち X_f の奇数番目の成分に注目したとき, $W^{\frac{N}{2}} = -1$ となる事を利用して,

$$X_{2m+1} = \sum_{n=1}^{\frac{N}{2}} x_n W^{(2m+1)n} + \sum_{n=\frac{N}{2}+1}^{N} x_n W^{(2m+1)n} \tag{8}$$

$$= \sum_{n=1}^{\frac{N}{2}} x_n W^{(2m+1)n} + \sum_{n=1}^{\frac{N}{2}} x_{n+\frac{N}{2}} W^{2mn} W^n W^{mN} W^{\frac{N}{2}} \tag{9}$$

$$= \sum_{n=1}^{\frac{N}{2}} (x_n - X_{n+\frac{N}{2}}) W^{2mn} W^n \tag{10}$$

と表わす事ができる。式7同様、

$$X_n^O \equiv (X_n - X_{n+\frac{N}{2}}) W^n \tag{11}$$

とおくと, 式10は $X_{2m+1}(m = 1, \cdots, \frac{N}{2})$ が長さ $\frac{N}{2}$ の数列 x_n^O のフーリエ変換となる事を表わしている。この分割を基数 R で割り切れる段階まで進めることで計算回数を減らしていく事ができる。$N = R^M$ で表わせる場合が最も効率良く, あるひとつの X についての和の計算が $\log N = M$ 回, 数列 X すべてについては,$N \log N$ 回の計算で結果を得る事ができる。

4.3 α の求め方

まず求めたパワースペクトルのパワー、周波数をそれぞれ対数で表す。そして対数で表した結果の低周波数部分を線形回帰する。線形回帰した結果の直線の傾きに-1をかけたものを α とする (図5)。

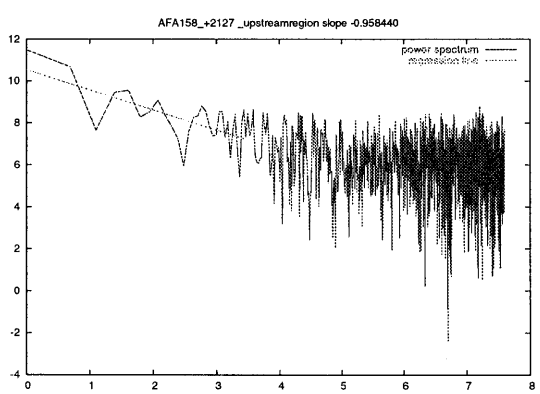

図 5. パワースペクトルとそこから求めた回帰直線

5 実験

Dictyostelium discoideum cDNA project[12] はキイロタマホコリカビという粘菌の一種の cDNA データを扱っているサイトである。キイロタマホコリカビの DNA 解析により得られた完全長 cDNA および、DictyosteliumBlast Search によるアライメントの結果などを公開している。このキイロタマホコリカビは細胞性粘菌類の中では世界中でもっとも研究対象とされているものである。

本実験で使用するデータは Dictyostelium discoideum cDNA project のデータをもとに筑波大学電子情報工学系安永研究室で作成された、粘菌転写開始点上流非転写領域ライブラリを使用した。これはそのアライメントの結果から転写開始点の位置を含む DNA を調べて、その上流を非転写領域として収集したものである。転写開始点の上流非転写領域を 2000base、転写開始点後を 100base と定めており、規定の 2100base が揃わない配列は、転写開始点が 2001base となるように位置を合わせて空欄は*で埋めてある (図6)。

図 6. AFA407_+1190 の FASTA 形式のデータ

5.1 α の導出

まず、粘菌非転写領域ライブラリの全てのデータに関して 2001 番目以後の転写領域とそれ以前の非転写領域に分ける。非転写領域のデータは 2000 番目から、1 番目方向にひっくり返した (図7)。

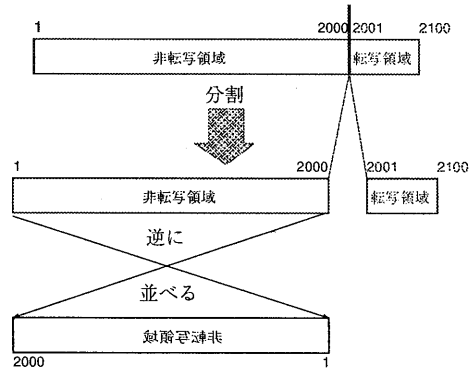

図 7. データの前処理

次に各塩基を以下のように数値に変換する。

(1) A → 0.75

(2) C → 0.25

(3) G → -0.25

(4) T → -0.75

数値に変換すると図8のようになる。

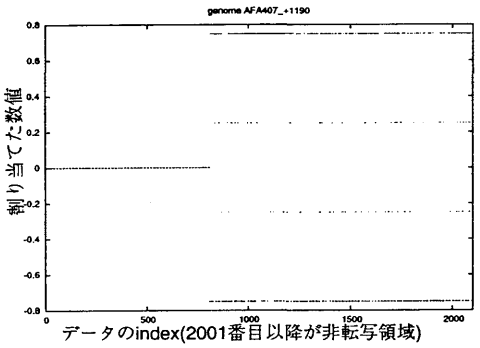

図 8. 数値に変換した AFA407_+1190 のデータ

そして、数値に変換した転写領域と非転写領域のデータのパワースペクトルを求める。その結果からそれぞれの領域の α を求める (4.3 節参照)。以上の操作を行った。一例を図9に示す。

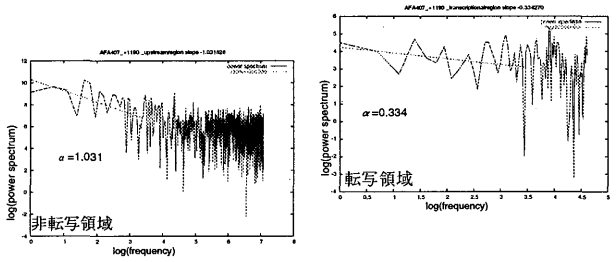

図 9. 転写領域と非転写領域の $1/f^{\alpha}$

5.2 転写領域と非転写領域の比較

その後それを集計し、各発現期毎に非転写領域、転写領域の α の平均を求めた結果が表1の通りである。

表 1. 各発現期の非転写領域と転写領域の $1/f^{\alpha}$

	非転写領域 α 平均	転写領域 α 平均
AF	0.449	0.192
CF	0.422	0.112
SF	0.423	-0.233
VF	0.593	0.063

表1のどの発現期においても非転写領域の方が α が大きい。このことから非転写領域は転写領域より規則性が高いように見える。各発現期毎の違いについては表1の結果からは特に読み取れない。Sfにおいて転写領域の値がマイナスになっているが、これはSfのデータを一つずつみていくと、とても短いデータの場合に1／fが成り立たずにマイナスの大きな値が出ていたためと思われる。

5.3 非転写領域の一定の長さの部分データの $1/f^{\alpha}$ の指数 α の変化

次の実験はある部分配列の長さLを決めた時、長さNのデータの場合1からL、2からL+1,…, N-L+1からNの合計N-L+1個の部分データ (図10) の $1/f^{\alpha}$ の指数 α を求める。一つのデータファイルに関してN-L+1個の部分配列の α が出る。

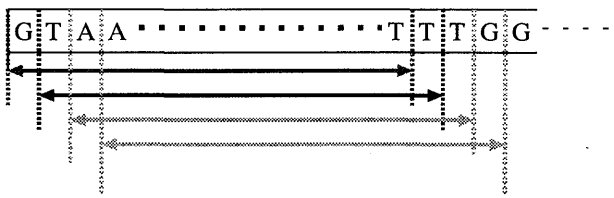

図 10. 部分配列の取り方

本実験では $L = 48$ とした。データの切り出し方以外は先の実験と同じである。結果はあるデータ AFA158_+2127 では図11のようになった。

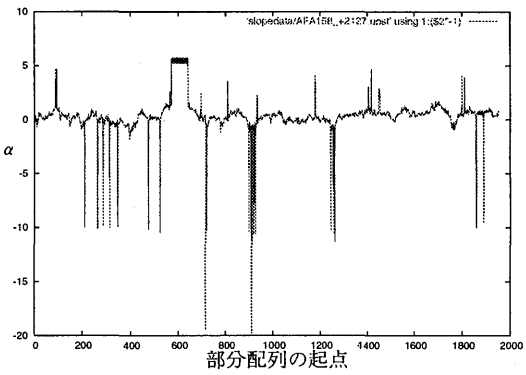

図 11. AFA158_+2127 の非転写領域の各部の α

当初これらのデータ群を分析することによって各発現期毎の違い等を見つけようとしたが、結局見つけ出すことができなかった。しかし、この試行錯誤のなかで興味深い特徴を見つけた。図11において、急激に上に凸となっている部分で、あるパターンが繰り返し続いていたり同じデータが続いているのを見つけることができる。そしてこれは他のデータに対しても同様の性質を見出すことができた。

6 おわりに

本研究では、細胞性粘菌のDNA塩基配列を数値化しパワースペクトルを求め、その低周波部分の傾きから$1/f^{\alpha}$揺らぎの指数αを算出し、それを不規則性の指標として分析を行った。

(1) 各発現期のαを比較し、どの発現期においても非転写領域の方が$1/f^{\alpha}$の指数部αが大きいことがわかった。これにより非転写領域の方がより規則性が高いように見える。ただし転写領域と非転写領域とでは含まれるA,G,C,Tの比率が違い、転写領域はA,G,C,Tがそれぞれ約25%だが、非転写領域の塩基配列の約80%がAとTで構成されている。そのため見かけ上規則性が高くなっているように見えると考えられる。

(2) 非転写領域において、あるデータ長の部分列を全区間にわたり全て取り出し、位置によってαがどのように変わるかを調べた。その結果αが急増している部分があり、その部分では塩基配列の繰り返しパターンが出現していることが分かった。このことから本手法はDNA塩基配列に含まれる繰り返しパターンの検出に役立つと思われる。

今後の課題としては、引き続き各発現期間の不規則の観点から見た特徴の調査、今回見つけた手法の実用性の検証、下に凸部分のデータにも共通点があるか分析すること等が挙げられる。

謝辞

本研究の一部は文部科学省科学研究費補助金(特定領域研究14015206)により行われた。関係各位に感謝する。

参考文献

[1] 中村桂子, 藤山秋佐夫, 松原謙一監訳. 細胞の分子生物学. Newton Press, 1995.

[2] 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.

[3] W. Li and K. Kaneko. Long-range correlation and partial $1/f^{\alpha}$ spactrum in a noncoding dna sequence. Europhys. Lett., Vol. 17, pp. 665–660, 1992.

[4] W. Li, K. Kaneko, P. J. Munson, R. C. Taylor, and G. S. Michaels. Dna correlations. Nature, Vol. 360, pp. 635–636, 1992.

[5] R. F. Voss. Evolution of long-range fractal correlation and 1/f noise in dna base sequence. Phys. Rev. Lett., Vol. 68, pp. 3805–3808, 1992.

[6] 澤岨英生, 宮城拓. バクテリオファージϕ-174 : Dna塩基配列のパワースペクトルにおけるフラクタル的充填. 琉球大学理学部紀要, Vol. 70, pp. 43–46, 2000.

[7] E. Takushi and H. Miyagi. Fractal packing of the dna sequence of bacteriophage ϕ-x174(ii). Bull. Facul. Sci.,Univ. Ryukyus, Vol. 71, pp. 21–23, 2001.

[8] E. Takushi and H. Miyagi. Fractal packing of the dna sequence of bacteriophage ϕ-x174(iii). Bull. Facul. Sci.,Univ. Ryukyus, Vol. 72, pp. 43–47, 2001.

[9] http://web.wtez.net/n/s/ns54007/gene/initiation.html.

[10] http://www.nias.affrc.go.jp/project/inegenome/press/c

[11] 田原孝. カオスと健康. からだの科学. 日本評論社, 1991.

[12] http://www.csm.biol.tsukuba.ac.jp/cdnaproject.html.