長距離相関を持った記号列の統計的解析

岩手大学人文社会科学部
五味壮平

1 初めに

高等動物のDNA塩基配列は、アミノ酸をコードし遺伝子として働く領域と、それ以外の領域(以下、非コード領域と呼ぶ)に分類される。非コード領域の一部は、遺伝子の活性などを調節する役割を担うことが知られているが、それ以外の部分については、遺伝・発生に関して何らかの意味があるのか、あるいは進化の過程で生じた単なるゴミであるのか、今のところ明らかにされていない。尚ヒトの場合、ゲノム全体で10^9塩基対程度の長さの塩基配列を持つが、このうち非コード領域は実に95 %を占めるといわれている。

近年、この非コード領域の塩基配列に統計的に奇妙な性質を持つことが発見された[1-24]。A(アデニン),G(グアニン),T(チミン),C(シトシン)の4文字が一次元的に並んだ配列に、長距離的な相関が存在するというのである。すなわち、非コード領域上で遠くはなれた二つの塩基の出現は独立ではないことになる。この事実は、相関関数のスペクトル表示が$1/f^{\alpha}$型($\alpha \sim 1$)になることや[2]、"DNA walk"と名付けられた解析[3]などにより示されている[1]。さらに、テキスト中に現れる単語についての経験則・Zipf則が、DNAの非コード領域の塩基配列に対しても成り立つことを主張する解析もなされるに至った[15,19 ~ 23] [2]。

このDNAの性質を説明することを念頭においたモデルが、多数提出されている[25-31]。これらは、長距離相関が進化の過程で生じた突然変異(重複や挿入など)の自然な帰結であると考えるモデル[25]から、あらかじめテキストを意識して"単語"を定義した上で、ある"文法"に従って単語の並びを決めていくというモデル[28]まで多岐にわたっている。これらのモデルの多くは、確かに長距離相関を持った記号(ビット)列を生み出す。

こうした状況にあって、もっとも問題だと思われるのは、スペクトル解析やべき的に減衰する相関関数の指数の測定、あるいはZipf則の成立などを調べる以外に、長距離相関を特徴づけるための有効な方法を我々があまり持ち合わせていないということである。当然、DNAの長距離相関を説明するためのモデルとしてどれが適当であるかを議論することはできない。またそもそも、これら異なるモデルから生み出されるビット列は統計的に区別し得るのかどうか、言い替えれば、モデルの区別が原理的に可能であるかどうかすら明らかではないのである。この小文では、提出されているモデルを主な題材とし、長距離相関を持つ記号列をいかにして特徴づけて行けばいいのか、その一般的方法について考察する[3]。

以上問題の背景として、DNAの長距離相関というトピックを引き合いに出した。この性質がどんなメカニズムにより生み出され、どのような意味を持つのかという問題は確

[1]以下、長距離相関を持つ記号列とは、こうした相関関数(もしくはそれと類似の物理量)の性質を持つものとする。

[2]これらの解析をもとに、非コード領域がテキストと同じような形で情報を蓄えているのではないかという過激な説もほのめかされている。(文字列に長距離的相関が成り立つという主張は、物語などのテキストにおいてもなされている。)

[3]この文は、"S.Gomi and K.Shindo, Phys.Rev.E (投稿中)"の内容を一部修正・加筆・削除したものである。

かに興味深い。しかし、長距離相関を持つ記号列を特徴づけるという研究は、例えば情報論的な観点からも重要であろう。また、長いテキストがやはり長距離相関を持つと報告されていることは、人間の思考や言語現象についても、そうした研究が重要な意味を持つ可能性を示唆している。例えば、脳内における情報の表現とその動作原理を、生み出された記号列が反映していることも考えられなくはない。以下の研究は、DNA の長距離相関という性質だけでなく、こうした問題意識からも動機づけられている。

この文章は、以下の構成を持つ。2 章では、3 章で扱う題材として、ビット列を生成する 2 つのモデルを簡単に紹介する。3 章では、ビット列中の短い配列についての確率分布の測定を、「くりこみ」と組み合わせて行う方法について述べる。またこの方法を、DNA の塩基配列に対して適用した結果を示す。最後の章は簡単なまとめにあてる。

2 モデル

2.1 伸長-修正システム (Expansion-Modification system)[25]

このモデル (以下 EMS と略) に属するものとして様々なものを考えることができるが、ここではそのうち最も単純なものを扱う。最初に"0"と"1"を等確率で選ぶ。このビットを次のルールに従って書き換える。

- "0"のビットは、確率 $p(0 < p < 0.5)$ で 1 に、$1-p$ の確率で"00"に書き換える。
- "1"のビットは、確率 p で 0 に、$1-p$ の確率で"11"に書き換える。

このルールを適用することにより生じたビット列に対して、更に同じルールを適用する。以下、同様なプロセスを繰り返すことにより、原理的にはどんな長さよりも長いビット列を得ることができる。そのビット列が"0"と"1"をやはり等確率で含むことは明らかであろう。このモデルは長距離相関を持ったビット列を生み出すわけであるが、パラメータ p は、相関の強さ (例えばスペクトル解析を行なった際の指数αなどで表される) に影響を与える。p が小さい程、相関の度合は強い (αは小さい)。

2.2 一般化された Lévy ウォークモデル (Generalized Lévy Walk Model)[27]

このモデル (以下 GLWM と略) では、(a)"1"を確率 $(1+\epsilon)/2$、"0"を確率 $(1-\epsilon)/2$ $(0 \leq \epsilon \leq 0.5)$ で含むブロック (複数のビットからなる) と (b)"0"を確率 $(1+\epsilon)/2$、"1"を確率 $(1-\epsilon)/2$ で含むブロック とを交互に配置することによりビット列を構成する。それぞれのブロックの長さ lは、次の分布関数 $D(l)$ によって確率的に決定されるものとする。

$$D(l) \sim l^{-\mu} \quad (l_{\min} < l (< l_{\max})), \tag{1}$$

ここで、$2 \leq \mu \leq 3$ は相関の強さを決めるパラメータであり、μが小さい程相関は強くなる。また $l_{\min}$及び $l_{\max}$は、それぞれ分布の上限と下限を与えるパラメータである。

3 短配列の出現確率分布による解析

一般に、定常な確率過程に関するあらゆる情報は、各事象がどんな確率で生起するか、すなわち事象の確率分布から得られる。従って、確率的な現象やモデルを比較する際に、

その確率分布を直接用いて議論するのは自然であろう。ただし今の問題の場合、モデル間で本来比較すべきなのは、「無限長のビット列の生起確率分布」であり、これを測定することはとてもできない。

しかし、ここで問題としているのは、「長距離相関を持つ」という、注釈つきのビット列を生み出すモデルである。相関関数やそのスペクトル表示がべき的な関数で記述されるという事実は、これらのモデルが何らかの形で自己相似性を持ち合わせていることを示唆する。そして、実際に自己相似性を持っているとすれば、非常に大きなスケールの性質、すなわち非常に長いビット列の統計的性質が、短いビット列の統計的性質にも反映しているとの期待を抱くことができる。そこで、以下では**短い配列の出現確率分布**を用いた特徴づけを試みる。具体的には 10 ビットの配列を考える。従って、$2^{10} = 1024$ 通りのヴァリエーションが、各モデルから生み出されたビット列中に、それぞれどのような確率で出現するかを測定するわけである[4]。

まず、モデルから生み出されたビット列をそのまま解析の対象とする。図 1 では、EMS、GLWM のそれぞれに対して、横軸に配列のヴァリエーション (10 進表示してある)、縦軸にその出現確率をプロットした。これらはそれぞれ特徴的な形をしており、しかも明確に異なっている。このことから短配列の出現確率分布がモデルの特徴づけに有効であり得ることが示唆される。

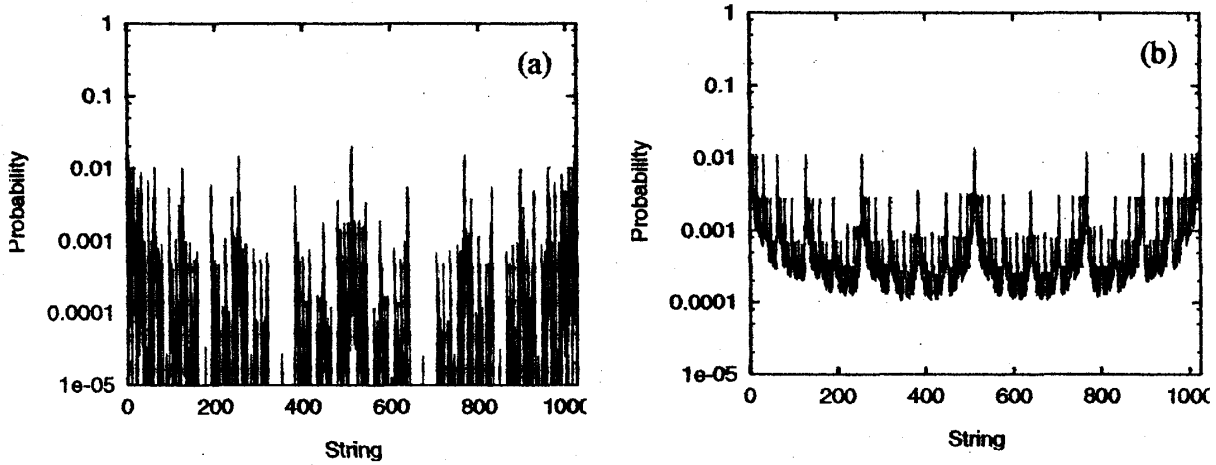

図 1: 短配列の出現確率分布 (a) EMS についての結果。$p = 0.1$ の場合。(b) GLWM についての結果。$\mu = 2.2, \epsilon = 0.6, l_{\min} = 10, l_{\max} = 10^9$の場合。いずれもモデルでも $N = 131072$ の長さのビット列を M=100 個用意し解析を行った。

しかし、上の議論にもあるように、ここでの解析は自己相似性という性質を基礎に据えたものであった。この性質をモデルが保有しない場合、短い配列についての確率分布は、単に局所的な特徴を反映したものに過ぎなくなる。そこで次に考えるべきことは、ビット列を「くりこ」むことである。ここで「くりこみ」とは以下の操作を指す。まず充分に長いビット列が与えられたとする。このビット列を d(d は奇数) ビットずつの領域に区切る。各々の領域の中に"1"より"0"の方が多ければ、この領域全体を一つの"0"に置き変える。また"0"より"1"の方が多ければ、全体を"1"に置き変える[5]。このくりこまれたビット列上で、10 ビットの短配列の出現確率分布を測定する。モデルが自己相似性を持つ場合には、こうして得られた確率分布は、もとのビット列についての分布と同形になることが期待さ

[4]各モデルを用いて充分に長いビット列 (長さを N とする) を多数 (その数を M とする) 生成し、これらのビット列上での短配列の出現頻度を数えることによって実際の測定を行った。

[5]従って、くりこみを行ったあと、ビット列の長さはもとの $1/d$ になる。

れる。

以下では $d = 11$ とし、各モデルから生成されたビット列にくりこみの操作を重ねて適用していく。そして、一回くりこむたびに短配列の出現分布を測定する。その結果を示すにあたって、図 3 のようなグラフでは分布の特徴をもう一つ掴むことができないので、別の形での表示を考える。表 1 ～ 2 では、出現確率が大きい配列を上位から各々24 個ずつ選んで並べている。これらは各モデルに対し、もとのビット列、およびくりこみを 1~3 回適用した際の分布から得られたものである。尚、これらの配列が作り出す"パターン"を見易くするために、配列中"1"のビットは●で、"0"のビットは○で表示してある。

期待に反して、各モデルともくりこんで行ったときに得られる分布は、もとのビット列についての分布とは異なっている。しかしよく見ると、くりこみの回数を増やして行くと、ほぼ安定な分布が出現してくることがわかる。すなわちこれらのモデルは比較的大きなスケールでの自己相似性を備えており、くりこみで局所的な性質が除かれることにより、それらが見えてきたものと考えられる [29]。そして、くりこんだ後のビット列に対して現れてくる安定な分布は、極めて特徴的である。まず EMS についての表は、(1) すべて"0"もしくは"1"である配列、(2) "0"("1") の中に 1 個、あるいは少数個の"1"("0") が浮かんでいる配列の二種類から成り立っている (表 1(b)-(d))。このタイプの分布を"タイプ A"と呼ぶことにしよう。一方 GLWM では、(1) すべて"0"もしくは"1"である配列、(2) 連続した"0"と連続した"1"の領域からなる配列の二種類が見られる (表 2(c)-(d))。これを"タイプ B"の分布と呼ぶ。これらのタイプの分布がなぜ得られるかは、以下のように理解される。

EMS で、例えば最初に"0"というビットから出発したとしよう。アルゴリズムに従ってビット列を成長させて行く際、p の値が小さければ、"0"の濃度は大きく保たれたままであろう。しかし、途中時折"0"から"1"への反転が起こる。反転が起きたビットからは、今度は"1"の濃度が高い領域が成長して行く。そして、さらにこの"1"が多い領域の中に"0"を多く含む部分的な領域が生まれてくる。重要なのはこうした現象があらゆるスケールで生ずることであり、これが自己相似的な性質を生み出すもとになっているのである。タイプ A の分布で、すべて"0"または"1"の配列、および"0"("1") の海の中に"1"("0") の島が浮かんでいるような配列が多く見られるのは、以上のような成長過程の反映である。

一方 GLWM で生み出されるビット列は、"0"が多く含まれるブロックと"1"が多く含まれるブロックから構成される。これらのブロックの長さはべき的に分布するので、非常に大きなブロックも出現し得る。つまり、どれだけくりこみを施しても"0"が連続する領域と"1"が連続する領域は残るわけである。こうした領域が存在することの反映として、すべて"0"("1") である配列がタイプ B の分布で最も多く現れる。また次に多く見られるのは、連続した"0"の領域と"1"の領域との境界から得られる配列 (すなわち連続した"0"と"1"からなる配列) なのである。

以上、この章では短配列の確率分布の測定を、くりこみという操作と組み合わせて行うことでモデルの特徴づけを試みた。重要なのは、くりこみを繰り返して安定な分布が現れた際、この分布が、ビット列の生成機構を見えやすい形で反映していたことである。このことは、例えば出所のわからないビット列と出会った時にも、その生成され方について、この解析法がヒントを与えてくれる可能性があることを意味している。

表 1: EMS における短配列の出現確率分布　出現確率の高い配列を 24 個選んできたもの。表中の○は"0"を●は"1"を表す。それぞれの配列の出現確率とその順位も示した。モデルのパラメータの値は $p = 0.1$ であり、くりこみの"倍率"は d=11 である。くりこみをしないときのビット列全体の長さは N= 131072、また用意したビット列の数は M=100 個である。一回くりこみを行うごとに、ビット列の数を 10 倍に増やして解析した。

表 1(a) EMS:くりこみ無し

配列	出現確率	出現順位
○○○○○○○○○○	0.176772	1
○○○○○○○○○●	0.018562	4
○○○○○○○○●○	0.008792	24
○○○○○○○○●●	0.015313	7
○○○○○○○●○○	0.008811	21
○○○○○○○●●●	0.009763	15
○○○○○○●●●●	0.010534	11
○○○○○●○○○○	0.008803	22
○○○○●○○○○○	0.008812	20
○○○○●●●●●●	0.010305	14
○○○●○○○○○○	0.008797	23
○○○●●●●●●●	0.009467	18
○○●○○○○○○○	0.008819	19
○○●●●●●●●●	0.014677	10
○●●●●●●●●●	0.017484	5
●○○○○○○○○○	0.018562	3
●●○○○○○○○○	0.015313	8
●●●○○○○○○○	0.009733	16
●●●●○○○○○○	0.010504	12
●●●●●○○○○	0.010313	13
●●●●●●○○○	0.009480	17
●●●●●●●○○	0.014685	9
●●●●●●●●○	0.017483	6
●●●●●●●●●	0.164886	2

表 1(b) EMS:くりこみ一回

配列	出現確率	出現順位
○○○○○○○○○○	0.098538	2
○○○○○○○○○●	0.017407	6
○○○○○○○○●○	0.010628	10
○○○○○○○●○○	0.010429	22
○○○○○○●○○○	0.010482	19
○○○○○●○○○○	0.010495	17
○○○○●○○○○○	0.010487	18
○○○●○○○○○○	0.010476	20
○○●○○○○○○○	0.010430	21
○○●●●●●●●●	0.010028	23
○●○○○○○○○○	0.010633	9
○●●●●●●●●●	0.017600	3
●○○○○○○○○○	0.017408	5
●○●●●●●●●●	0.010725	8
●●○●●●●●●●	0.010505	16
●●●○●●●●●●	0.010560	14
●●●●○●●●●●	0.010597	13
●●●●●○●●●●	0.010600	12
●●●●●●○●●●	0.010602	11
●●●●●●●○●●	0.010559	15
●●●●●●●●○○	0.010012	24
●●●●●●●●○●	0.010742	7
●●●●●●●●●○	0.017600	4
●●●●●●●●●●	0.099561	1

表 1(c) EMS:くりこみ二回

配列	出現確率	出現順位
○○○○○○○○○○	0.085459	1
○○○○○○○○○●	0.016586	4
○○○○○○○○●○	0.010687	7
○○○○○○○○●●	0.009089	24
○○○○○○○●○○	0.010579	13
○○○○○○●○○○	0.010647	9
○○○○○●○○○○	0.010514	16
○○○○●○○○○○	0.010471	19
○○○●○○○○○○	0.010617	10
○○●○○○○○○○	0.010527	15
○●○○○○○○○○	0.010652	8
○●●●●●●●●●	0.016417	6
●○○○○○○○○○	0.016588	3
●○●●●●●●●●	0.010593	12
●●○○○○○○○○	0.009128	23
●●○●●●●●●●	0.010473	18
●●●○●●●●●●	0.010531	14
●●●●○●●●●●	0.010400	21
●●●●●○●●●●	0.010393	22
●●●●●●○●●●	0.010487	17
●●●●●●●○●●	0.010460	20
●●●●●●●●○●	0.010599	11
●●●●●●●●●○	0.016421	5
●●●●●●●●●●	0.084967	2

表 1(d) EMS:くりこみ三回

配列	出現確率	出現順位
○○○○○○○○○○	0.084409	1
○○○○○○○○○●	0.016228	4
○○○○○○○○●○	0.010857	7
○○○○○○○○●●	0.008862	24
○○○○○○○●○○	0.010703	12
○○○○○○●○○○	0.010582	20
○○○○○●○○○○	0.010619	16
○○○○●○○○○○	0.010646	15
○○○●○○○○○○	0.010592	18
○○●○○○○○○○	0.010712	11
○●○○○○○○○○	0.010812	8
○●●●●●●●●●	0.016161	5
●○○○○○○○○○	0.016241	3
●○●●●●●●●●	0.010768	9
●●○●●●●●●●	0.010683	13
●●●○●●●●●●	0.010545	22
●●●●○●●●●●	0.010603	17
●●●●●○●●●●	0.010591	19
●●●●●●○●●●	0.010563	21
●●●●●●●○●●	0.010661	14
●●●●●●●●○○	0.008886	23
●●●●●●●●○●	0.010758	10
●●●●●●●●●○	0.016149	6
●●●●●●●●●●	0.083798	2

表 2: GLWM における短配列の出現確率分布　表 1 と同様。 ただしモデルのパラメータは、$\mu = 2.2$, $\epsilon = 0.6$, $l_{\min} = 10$, and $l_{\max} = 10^9$, である。

表 2(a) GLWM: くりこみ無し

配列	出現確率	出現順位
○○○○○○○○○○	0.043857	2
○○○○○○○○○●	0.012256	6
○○○○○○○○●○	0.011260	15
○○○○○○○●○○	0.011032	18
○○○○○○●○○○	0.010980	20
○○○○○●○○○○	0.010954	22
○○○○●○○○○○	0.010962	21
○○○●○○○○○○	0.010989	19
○○●○○○○○○○	0.011039	17
○○●●●●●●●●	0.004510	23
○●○○○○○○○○	0.011255	16
○●●●●●●●●●	0.012780	4
●○○○○○○○○○	0.012257	5
●○●●●●●●●●	0.011791	8
●●○●●●●●●●	0.011572	9
●●●○●●●●●●	0.011521	12
●●●●○●●●●●	0.011494	14
●●●●●○●●●●	0.011536	11
●●●●●●○●●●	0.011515	13
●●●●●●●○●●	0.011554	10
●●●●●●●●○○	0.004507	24
●●●●●●●●○●	0.011795	7
●●●●●●●●●○	0.012780	3
●●●●●●●●●●	0.046132	1

表 2(b) GLWM: くりこみ一回

配列	出現確率	出現順位
○○○○○○○○○○	0.200958	1
○○○○○○○○○●	0.009326	6
○○○○○○○○●○	0.005736	9
○○○○○○○○●●	0.004986	12
○○○○○○○●○○	0.004842	18
○○○○○○●○○○	0.004045	22
○○○●○○○○○○	0.004053	21
○○●○○○○○○○	0.004847	17
○○●●●●●●●●	0.004943	14
○●○○○○○○○○	0.005744	8
○●●●●●●●●●	0.009344	4
●○○○○○○○○○	0.009344	3
●○●●●●●●●●	0.005765	7
●●○○○○○○○○	0.004994	11
●●○●●●●●●●	0.004881	15
●●●○●●●●●●	0.004089	19
●●●●○●●●●●	0.003868	23
●●●●●○●●●●	0.003863	24
●●●●●●○●●●	0.004065	20
●●●●●●●○●●	0.004848	16
●●●●●●●●○○	0.004976	13
●●●●●●●●○●	0.005716	10
●●●●●●●●●○	0.009330	5
●●●●●●●●●●	0.200941	2

表 2(c) GLWM: くりこみ二回

配列	出現確率	出現順位
○○○○○○○○○○	0.123604	2
○○○○○○○○○●	0.008383	6
○○○○○○○○●○	0.004170	23
○○○○○○○○●●	0.006305	10
○○○○○○○●●●	0.005128	14
○○○○○○●●●●	0.004565	18
○○○○○●●●●●	0.004396	20
○○○○●●●●●●	0.004590	16
○○○●●●●●●●	0.005189	11
○○●●●●●●●●	0.006353	8
○●○○○○○○○○	0.004226	21
○●●●●●●●●●	0.008497	4
●○○○○○○○○○	0.008501	3
●○●●●●●●●●	0.004226	22
●●○○○○○○○○	0.006359	7
●●●○○○○○○○	0.005174	12
●●●●○○○○○○	0.004587	17
●●●●●○○○○○	0.004416	19
●●●●●●○○○○	0.004603	15
●●●●●●●○○○	0.005171	13
●●●●●●●●○○	0.006324	9
●●●●●●●●○●	0.004151	24
●●●●●●●●●○	0.008392	5
●●●●●●●●●●	0.125499	1

表 2(d) GLWM: くりこみ三回

配列	出現確率	出現順位
○○○○○○○○○○	0.046360	2
○○○○○○○○○●	0.005775	5
○○○○○○○○●○	0.003406	17
○○○○○○○○●●	0.004370	9
○○○○○○○●●●	0.003598	16
○○○○○○●●●●	0.003191	22
○○○○○●●●●●	0.003111	24
○○○○●●●●●●	0.003295	20
○○○●●●●●●●	0.003764	12
○○●●●●●●●●	0.004672	8
○●○○○○○○○○	0.003618	14
○●●●●●●●●●	0.006285	4
●○○○○○○○○○	0.006324	3
●○●●●●●●●●	0.003608	15
●●○○○○○○○○	0.004711	7
●●●○○○○○○○	0.003788	11
●●●●○○○○○○	0.003302	19
●●●●●○○○○○	0.003141	23
●●●●●●○○○○	0.003237	21
●●●●●●●○○○	0.003626	13
●●●●●●●●○○	0.004367	10
●●●●●●●●○●	0.003360	18
●●●●●●●●●○	0.005743	6
●●●●●●●●●●	0.046654	1

表 3: 哺乳類 DNA における短配列の出現確率分布　哺乳類 DNA 塩基配列 19 本をビット列に変換 (アデニンとグアニンを"1"、シトシンとチミンを"0"に変換) し、表 1 ～ 2 と同様に短配列の出現確率分布を求めた。これらの DNA の塩基配列のデータは、GenBank Release No. 94.0. から得られたもので、そのアクセスコードと塩基配列の長さは以下の通りである。HSABLGR2 (59012 塩基対 (bp)), HSABLGR3 (84539bp), HSG6PDGEN (52173bp), HSMHCAPG (66109bp), HSTCRBV (77743bp), HSU07000 (152141bp), HUMFMR1S (152351bp), HUMGHCSA (66495bp), HUMHBB (73308bp), HUMHDABCD (58864bp), HUMHPRTB (56737bp), HUMMMDBC (68468bp), HUMNEUROF (100849bp), HUMRETBLAS (180388bp), HUMTCRADCV (97634bp), HUMTCRB (684973bp), HUMVITDBP (55136bp), MMBGCXD (55856bp), MUSTCRA (94647bp)。

表 3(a) DNA:くりこみ無し

配列	出現確率	出現順位
○○○○○○○○○○	0.008715	1
○○○○○○○○○●	0.003256	3
○○○○○○○○●○	0.002563	17
○○○○○○○○●●	0.002426	21
○○○○○○○●○○	0.002592	14
○○○○○○●○○○	0.002703	12
○○○○○●○○○○	0.003046	8
○○○○●○○○○○	0.003101	5
○○○●○○○○○○	0.002747	11
○○●○○○○○○○	0.002566	16
○○●●●●●●●●	0.002384	24
○●○○○○○○○○	0.002664	13
○●●●●●●●●●	0.003094	6
●○○○○○○○○○	0.003256	4
●○●●●●●●●●	0.002394	23
●●○●●●●●●●	0.002474	20
●●●○●●●●●●	0.002541	18
●●●●○●●●●●	0.002770	10
●●●●●○●●●●	0.002917	9
●●●●●●○●●●	0.002578	15
●●●●●●●○●●	0.002415	22
●●●●●●●●○●	0.002496	19
●●●●●●●●●○	0.003094	7
●●●●●●●●●●	0.008230	2

表 3(b) DNA:くりこみ一回

配列	出現確率	出現順位
○○○○○○○○○○	0.010206	1
○○○○○○○○○●	0.003715	6
○○○○○○○○●○	0.003381	15
○○○○○○○●○○	0.003381	16
○○○○○●○○○	0.003582	9
○○○○○●●○○	0.002701	23
○○○○○●○○○○	0.003474	12
○○○○○●●○○○	0.002588	24
○○○○●○○○○○	0.003272	21
○○○●○○○○○○	0.003592	8
○○●○○○○○○○	0.003282	20
○●○○○○○○○○	0.003287	19
○●●●●●●●●●	0.003720	4
●○○○○○○○○○	0.003715	7
●○●●●●●●●●	0.003346	17
●●○●●●●●●●	0.003415	13
●●●○●●●●●●	0.003779	3
●●●●○●●●●●	0.003395	14
●●●●●○●●●●	0.003194	22
●●●●●●○●●●	0.003533	10
●●●●●●●○●●	0.003484	11
●●●●●●●●○●	0.003336	18
●●●●●●●●●○	0.003720	5
●●●●●●●●●●	0.009133	2

表 3(c) DNA:くりこみ二回

配列	出現確率	出現順位
○○○○○○○○○○	0.025343	2
○○○○○○○○○●	0.006663	3
○○○○○○○○●○	0.004151	22
○○○○○○○○●●	0.005626	7
○○○○○○○●○○	0.005025	9
○○○○○○○●●●	0.004588	14
○○○○○●○○○	0.004479	17
○○○○○●○○○○	0.004752	12
○○○○●○○○○○	0.004697	13
○○○●○○○○○○	0.004588	15
○○●○○○○○○○	0.005134	8
○○●●●●●●●●	0.004151	21
○●○○○○○○○○	0.004970	10
○●●●●●●●●●	0.005844	5
●○○○○○○○○○	0.006663	4
●○●●●●●●●●	0.004042	24
●●○○○○○○○○	0.004861	11
●●○●●●●●●●	0.004096	23
●●●○●●●●●●	0.004479	16
●●●●○●●●●●	0.004315	19
●●●●●●●○●●	0.004260	20
●●●●●●●●○●	0.004369	18
●●●●●●●●●○	0.005790	6
●●●●●●●●●●	0.025834	1

ここで、応用例として、以上の方法をデータベースから取り寄せた哺乳類、主にヒトのDNA塩基配列に対して適用した例を紹介しよう。配列中、アデニンとグアニンは"1"、シトシンとチミンは"0"に置き換えてビット列に変換し、短配列の出現確率分布を測定した結果が表3である[6]。この表でまず目を引くのが、くりこみを行うか行わないかに関わらず、これらの分布が極めて類似していることである。すなわち、DNA塩基配列は1塩基のスケールから、より大きいスケールまで通じる精巧な自己相似性を持っていることになる。さらに、ここで見られた安定な分布はタイプAのものであった。GLWMやそれと類似のモデル[30]は、少なくともこの事実に関する限り、DNA配列の構造を説明するモデルとしては適切なものではない。尚、以上は非コード領域を含むDNA配列の解析結果であった。我々は非コード領域をほとんど含まず、長距離相関を持たないと主張されている酵母 (saccharomyces cerevisiae) のDNAについても解析を行ったが、表3とほぼ同様な結果が得られた。最近の研究で、コード領域でも長距離的な相関があるという報告もなされており[30]、コード領域と非コード領域の塩基配列が極端に違う構造を持っているという考えは単純には肯定できない。

4 まとめ

この小文では、長距離相関をもつビット列、もしくはそれを生み出すモデルを特徴づける一般的な方法として、「くりこみ+短配列の出現確率分布」が有効であることを示した。この解析法の強みは、ビット列が如何にして生成されたかについても手がかりを掴めることにある。また短配列の代表的な分布として、タイプAおよびタイプBの二種類が得られたが、これ以外にどのようなタイプの分布が現れ得るのか、という問題はとても興味深い。これを考えることは今後の課題としたい。

参考文献

[1] DNAの長距離相関に関する文献一覧が、W. Leeによってホームページの形で公開されている。URLアドレスは"http://linkage.rockefeller.edu/wli/dna_corr/list.html" である。

[2] W. Li and K. Kaneko, Europhys. Lett. **17**, (7) 655 (1992).

[3] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Nature **356**, 168 (1992).

[4] W. Li and K. Kaneko, Nature **360**, 635 (1992).

[5] P. Munson, R.C. Taylor, and G.S. Michaels, Nature **360**, 636 (1992).

[6] S. Nee, Nature **357**, 450 (1992).

[7] R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[8] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, F. Sciortino, and H.E. Stanley, Phys. Rev. Lett. **71**, 1776 (1993).

[9] R.F. Voss, Phys. Rev. Lett. **71**, 1777 (1993).

[10] C.A. Chatzidimitriou-Dreismann, and D. Larhammar, Nature **361**, 212 (1993).

[6]尚、データ量の関係上くりこみは二回までが限度であった。

[11] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[12] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, and H.E. Stanley, Phys. Rev. E **47**, 3730 (1993).

[13] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[14] W. Li, T.G. Marr, and K. Kaneko, Physica D **75**, 392 (1994).

[15] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).

[16] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[17] M.Y. Azbel, Phys. Rev. Lett. **75**, 168 (1995).

[18] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).

[19] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **52**, 2939 (1995).

[20] N.E. Israeloff, M. Kagalenko, and K. Chan, Phys. Rev. Lett. **76**, 1976 (1996).

[21] S. Bonhoeffer, A.V.M. Herz, M.C. Boerlijst, S. Nee, M.A. Nowak, and R.M. May, Phys. Rev. Lett. **76**, 1977 (1996).

[22] R.F. Voss, Phys. Rev. Lett. **76**, 1978 (1996).

[23] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. Lett. **76**, 1979 (1996).

[24] Pedro Bernaola-Galván, Ramón Román-Roldán, and José L. Oliver Phys. Rev. E **53** 5181 (1996).

[25] W. Li, Phys. Rev. A **43**, 5240 (1991).

[26] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23**, (5) 373 (1993).

[27] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[28] I. Kanter and D.A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[29] A. Czirók, R.N. Mantegna, S. Havlin, and H.E. Stanley, Phys. Rev. E **52** 446 (1995).

[30] P. Allegrini, M. Barbi, P. Grigolini, and B.J. West Phys. Rev. E **52** 5281 (1995).

[31] M.S. Vieira and H.J. Herrmann, Europhys. Lett. **33**, (5) 409 (1996).