待ち行列理論の遺伝子工学への応用

Hiroshi Toyoizumi and Yasuo Tanioka *

平成 15 年 12 月 22 日

概 要

待ち行列理論の Lindley 方程式を BioInformatics に応用し、遺伝子配列の類似度を比較する方法について述べる。遺伝子情報の塩基配列解析により、生物学的な機能を調査・研究するさまざまな試みが BioInformatics として行われている。同一の生物学的機能を実現する塩基配列間の差は、世代間のデータ受け渡しエラーとして確率的にモデル化できることが予想される。文字をサンプル空間として持つ確率過程として塩基配列を取り扱い、その性質を分析することが生物学的に重要となる。本論文では、待ち行列理論を使って、塩基配列を分析する手法を提案する。待ち行列理論では、到着過程の周辺分布や自己相関などの統計的性質が異なる場合には、待ち時間の統計的性質は大きく異なることがよく知られている。ゲノム情報内の塩基配列を到着過程としてモデル化することにより、対応する待ち行列システムの「待ち時間」として、異なる遺伝子をコードした塩基配列の統計的性質の差、すなわち生物学的機能の差を定量的に把握する可能性を実際の塩基配列を例に検討する。

1 Introduction

生物の遺伝子情報は、染色体内に塩基配列の形でコーディングされている。2001 年のはじめ、人間の遺伝子情報を含む塩基配列の解析は、Venter らのセレーラ社のグループ [7] および国際コンソーシアムのヒトゲノムチーム (the Genome International Sequencing Consortium)[4] によって、ほぼ終了した。現段階では、人間の遺伝子は約 30,000 と言われている [8, 7, 4]。また、人間以外の生物についても、着々と塩基配列の解析が行われている。得られた塩基配列情報を使い、遺伝的な機能を調査・研究するさまざまな試みが BioInformatics として行われ、生物の持つ膨大な塩基配列情報の中から、遺伝的・機能的に有意味な部分を抜き出し、その生物学的な機能を明らかにすることが重要となっている [2, 13, 8, 6]。

ある塩基配列の生物学的な機能を知るためには、その塩基配列の機能を停止させた個体を観察することによって得られる。しかし、遺伝子機能の停止を観察する実験は、人間の場合は禁止的である。また、人間以外の動物を使った

*University of Aizu, Performance Evaluation Lab., Tsuruga, Ikki-machi, Aizu-wakamatsu, Fukushima, Japan 965-8580 E-mail: toyo@u-aizu.ac.jp.

場合の実験コストも膨大である。これを克服する方法は、すでに他の生物によって知られている遺伝子との類比によって、遺伝子の機能を推測することである。たとえば、人間の遺伝子情報に対応するマウスの遺伝子情報部分を見つけることができれば、実験等によって知られている既知の遺伝子の機能との類推によって、遺伝的な治療が人間にどのような影響を与えるのかを知ることが可能になる。実際に、このような目的のために作られた多数の BioInformatics 用のツールやデータベースが整備されている [2]。機能的に似た配列を発見する基礎技術が、「アラインメント」と呼ばれるものである [10, 6]。アラインメントは、未知の塩基配列の中から、機能が既知の塩基配列と「似ている」部分を抜き出す技術である。生物の種が異なる場合には、当然、コードされている配列は異なる。しかし、生体のとして同じような機能を実現している場合には、遺伝的に同一の「祖先」を持ち、両者の配列の違いは、世代間の配列データの受け渡しのエラーとして捕らえることが可能である。したがって、同一の生物学的機能を実現する塩基配列間の差は、確率的にモデル化できることが予想される。塩基配列を文字列をサンプル空間として持つ確率過程として取り扱うことが可能になる。すでに、塩基配列の性質を文字列の Markov 過程的な性質によって記述することが試みられている [6]。

確率過程は、その数学的性質の研究 [3, 14] とともに、さまざまな物理現象をモデル化することが研究されている。例えば、金融、生物、通信、コンピュータサイエンスなどさまざまな場面での応用が考えられている。特に、確率過程の応用研究が進んでいる分野として、システムの性能を評価するために用いられる待ち行列理論がある [15, 11]。待ち行列理論では、システムに到着する客の到着間隔やサービス時間の統計的性質を仮定することによって、システムの性能である待ち時間や系内客数の統計的性質が得られる。また、到着間隔列の周辺分布や自己相関などの統計的性質が異なる場合には、待ち時間の統計的性質は大きく異なることがよく知られている。

本論文では、待ち行列理論で知られた Lindley equation を用いて遺伝子情報をモデル化する方法を述べ、実際に生物の塩基配列を処理する。また、待ち行列理論における既知の結果を使い、処理されたデータの解析を試みる。

2　遺伝子情報とその進化的な基礎

生物の遺伝子情報は、細胞内にある染色体上に 4 種類の塩基配列の形でコードされている。これをメッセンジャー RNA が転写し、たんぱく質を合成する。異なるコードからは、異なるたんぱく質が生成される。これが、細胞の異なる機能の実現や種の違い、個体の違いに反映される。

染色体は DNA と呼ばれる巨大な分子よって構成されている。塩基（nucleodite) と呼ばれる分子が 2 重らせん構造をとって、結合することによって DNA を形作っている。DNA を構成する塩基は 4 種類あり、それぞれアデニン (A)、

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	2																						
R	-2	6																					
N	0	0	2																				
D	0	-1	2	4																			
C	-2	-4	-4	-5	12																		
Q	0	1	1	2	-5	4																	
E	0	-1	1	3	-5	2	4																
G	1	-3	0	1	-3	-1	0	5															
H	-1	2	2	1	-3	3	1	-2	6														
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5													
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6												
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5											
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6										
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9									
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6								
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2							
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3						
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17					
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10				
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4			
B	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2		
Z	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

表 1: PAM250 Matrix

チミン (T)、グアニン (G)、シトシン (C) と呼ばれる。これらの塩基が３つ一組となり、20 種類のアミノ酸および読み出し開始、終始のマークなどをコーディングしている。たとえば、「ATG」という配列は、アミノ酸の一種である「メチオニン（M)」に相当する、またこのメチオニンは RNA が転写を開始するマークにもなっている。DNA 上に３つの塩基としてコードされたアミノ酸の情報は、メッセンジャー RNA に転写され、アミノ酸を部品として作った巨大分子であるたんぱく質を作り出すのに利用される。以下では、塩基配列データを４種類の塩基のデータがコードされたものとしてではなく、２０種類のアミノ酸の情報をコードしたものとみなす。

地球上の生物のゲノムの情報は、突然変異などで変化しながら受け継がれてきている。親から子にゲノム情報が引き渡されるとき、DNA の複製が行われる。DNA の複製の過程で親と子の DNA は進化に中立な形のランダムな

変動を受け、異なる可能性がある。この変異が子の個体の機能の維持に悪い影響を与える場合には、その変異は将来の世代には受け継がれない可能性が高い（自然淘汰：Natural selection)。一方、変動が生命の機能の維持に影響を与えない場合には、その変動は遺伝子情報内に蓄積される。この場合には、DNA 上にコーディングされた塩基配列は、同一の祖先を持つ個体間で異なるが、生物学的には、同一の機能を保持している可能性が高い。実際、似たような化学的性質（疎水性や電荷、極性など）を持つアミノ酸をコーディングしている場合には、入れ替わりが起きても、同一の生物学的機能が保持される可能性が高い。アミノ酸の入れ替わりのしやすさを定量的にあらわしたものとして、BLOSUM や PAM [5] と呼ばれる行列に集約されている。表 1 に BioInformatics でよく使われるアミノ酸置換マトリックス PAM250 を示す。対応する数字大きいほど、入れ替わりやすいことを表している。例えば、アミノ酸 R と H の入れ替わりやすさは +2 であり、R と G は −3 である。したがって、生物のアミノ酸配列では R は H と入れ替わりやすいが、G とは入れ替わりにくいことがわかる。

3　アミノ酸配列の類似度

3.1　アミノ酸配列の例

アミノ酸の配列の例と類似度を石川・金久 [9] p.65 の例にならって示そう。以下に示す 6 本の文字列は、1 列目が未知の配列で、後の 5 本がそれぞれ別の種類のレトロウィルスのある酵素が持つアミノ酸の配列[1]である。今、この未知の配列 unknown が他の 5 つの配列と似たような遺伝子機能をコーディングしたものかを判断したいとする。

unknown : ILDFHEKLLHPGIQKTTKLFGETYYFPNSQLLIQNIINECSICNLAK

MMULV : LLDFLLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN

HTLV : LQLSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGN

RSV : YPLREAKDLHTALHIGPRALSKACNISMQQAREVVQTCPHCNSA

MMTV : IHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVAT

SMRV : LESAQESHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGS

アライメント処理を行い、同じアミノ酸や性質の似たアミノ酸が縦に同じ位置になるように、ギャップ (gap) を入れ調整する（下記参照)。縦方向に揃った文字で構成される配列を、そのアライメントのコンセンサス配列 (consensus

[1]レトロウィルスは自分の遺伝子情報を宿主の DNA に刷り込んで増殖するする。この例の酵素はエンドヌクレアーゼと呼ばれるもので、DNA を切る働きをする。また、例えば HTLV はヒト T 細胞白血病ウィルスである。

sequence) と呼ぶ。

unknown : IL-DF----HEKLLHPGIQKTTK-LF--GET-YY-FPNSQLLIQNIINECSICNL-AK
M − MULV : LL-DFL--LHQ-LTHLSFSKM-KALLERSHSPYYMLNRDRTL-KNITETCKACAQ-VN
HTLV : LQLSPA-ELHS-FTHCGQTAL-T-LQ--------GATTTEA--SNILRSCHACRG-GN
RSV : YPLREAKDLHT-ALHIGPRAL-S-KA-------CNISMQQA--REVVQTCPHC-N-SA
MMTV : IH-EAT-QAHT-LHHLNAHTL-R-LL-------YKITREQA--RDIVKACKQCVV-AT
SMRV : LE-SAQ-ESHA-LHHQNAAAL-R-FQ-------FHITREQA--REIVKLCPNCPDWGS
Consensus : ---------H----H---------------------------------C--C-----

ここでは、H で示されるヒスチジンというアミノ酸が 2 個と、C で示されるシステインが 2 個が縦に揃っており、コンセンサス配列とみることができる。また、コンセンサス配列のなかに見られるパターンが、アライメントされた配列群の遺伝的機能を特徴づけるものと判断できるとき、そのパターンを、配列モチーフ (sequencemotif) とか、単にモチーフと呼ぶ。C が 2 個と H が 2 個で特徴づけられるパターンは、「亜鉛の指」(ジンクフィンガー,zinc finger) と呼ばれる有名なモチーフであることがわかる。したがって、生物学的機能には、この unknown の配列は、他のレトロウィルスの酵素と同じ性質を持つであろうことが予想される。しかし、このような unknown の配列がまったくの偶然に他の 5 つの酵素の配列と同じ構造を持つ可能性もある。この可能性を定量的に評価するために確率論的な考え方が必要となる。

3.2 類似度のスコアリング

アミノ酸をあらわす 20 種類の文字で構成された長さ N の二つの文字列 $\{X_n\}$、$\{Y_n\}$ を考える。ここで、ベースとなる $\{X_n\}$ に対して、$\{Y_n\}$ がどれだけ「似ているか」を定量的に把握したいとする。Section 3.1 で見たように、各文字列中で、まったく同じ文字の部分あるが、違った部分もある。アミノ酸の中には同じような性質を持つものも多いので、これを考慮にいれて、類似度を考える必要がある。ここでは、[6] にしたがって、類似度をあらわすスコアを定義する。

Section 2 でもみたように、生物学的機能を表すアミノ酸配列は静的なものではなく、突然変異や自然淘汰などのランダムな影響を受けるために、種や個人の単位で変化することが知られている。したがって、生物学的機能を表すアミノ酸配列は、ランダムな変動を含む確率変数の列と考えることができる[2]。アミノ酸配列中の文字 X の周辺分布を $q(a) = P[X = a]$ とする。また、遺伝学上意味のある 2 つの異なるアミノ酸配列から得られた文字をそれ

[2]遺伝学上意味のあるアミノ酸配列の確率法則は、アミノ酸の化学的性質や既存の塩基配列データベースから得られる。

ぞれ X と X' とする。遺伝学上意味のある二つの配列は、進化の過程で共通の祖先を持つ場合などがあるため、一般には独立とは考えられない。この同時分布を $q(a,b) = P[(X,X') = (a,b)]$ とする。２つのアミノ酸が同じ機能を実現するには、同じアミノ酸である確率が高いので、$P[X = a \mid X' = b]$ より $P[X = a \mid X' = a]$ が大きいことが予想される。したがって、以下では下記の不等式が成立することを仮定する。

$$\frac{q(a,a)}{q(a)q(a)} \geq \frac{q(a,b)}{q(a)q(b)} \quad \text{for all } b. \tag{1}$$

また、同じような化学的性質を持つアミノ酸 a と a' の場合には、a と異なる化学的性質を持つアミノ酸 b と比較すると、入れ替わりが起きる可能性が高いので、

$$\frac{q(a,a')}{q(a)q(a')} \geq \frac{q(a,b)}{q(a)q(b)}. \tag{2}$$

が成立すると仮定する。ここで、すでに機能がわかっているアミノ酸配列中の文字 X と未知の配列中から得られた文字 Y の類似度をあらわす「スコア」を次のように定義する。

Definition 1 (文字ペアの類似度スコア). *2* つの文字がそれぞれ *a, b* であったとき、この文字ペアの類似度スコア $s(a,b)$ を次のように定義する。

$$s(a,b) = \log\left(\frac{q(a,b)}{q(a)q(b)}\right). \tag{3}$$

このように定義されたスコアは、そのアミノ酸を入れ替えた場合にも同一の生物学的機能を実現する場合には正となり、逆に、生物学的機能を破壊するような場合には、負となる。実際、アミノ酸 a と a' の科学的特性が類似しているとしよう。その場合には、a と a' が入れ替わるような突然変異が起こっても同じ生物学的機能を維持し、自然淘汰の影響を受けない可能性が高い。したがって、現在得られたアミノ酸の配列において、a とはまったく独立に a' が生じるよりも、a から a' に変化するようなケースが多いと考えられる。よって、$q(a,a') = P[(X,Y) = (a,b)] \geq P[X = a]P[Y = b]$ であり、スコアは $s(a,b) \geq 0$ である。一方、他のアミノ酸に変わると自然淘汰などの影響を受け、その変化が保存されにくい場合には、$P[(X,Y) = (a,b)] < P[X = a]P[Y = b]$ となり、スコアは $s(a,b) < 0$ となる。より一般に、スコアに対して、次の定理が得られる。

Theorem 1. *1.* 文字ペア X と Y が独立で、同一の分布 *q(x)* に従うとき、その類似度スコアの期待値は $E[S] \leq 0$ となる。

2. 任意の文字ペアをあらわす確率変数 X と Y について、

$$E[s(X,X)] \geq E[s(X,Y)]. \tag{4}$$

Proof. 1. X と Y の独立性より

$$\begin{aligned} E\left[\frac{q(X,Y)}{q(X)q(Y)}\right] &= \sum_{x,y}\frac{q(x,y)}{q(x)q(y)}P[X=x,Y=y] \\ &= \sum_{x,y}\frac{q(x,y)}{q(x)q(y)}q(x)q(y) \\ &= \sum_{x,y}q(x,y)=1. \end{aligned}$$

Jensen の不等式より任意の上に凸な関数について、$E[f(X)] \leq f(E[X])$ なので、f として log をとれば、

$$0 = \log\left(E\left[\frac{q(X,Y)}{q(X)q(Y)}\right]\right) \geq E\left[\log\left(\frac{q(X,Y)}{q(X)q(Y)}\right)\right] = E[S].$$

2. (1) より

$$\frac{q(X,X)}{q(X)q(X)} \geq \frac{q(X,Y)}{q(X)q(Y)}.$$

両辺の log をとって期待値をとれば、

$$E[s(X,X)] \geq E[s(X,Y)].$$

□

PCR 法やショットガン法によって得た塩基配列やアミノ酸配列の中には、生物学的機能の遺伝子をコードしている部分とそうでない部分（ジャンク）が混じっていることが知られている。ジャンクの部分は、生物学的な機能と無関係なため、ランダムにアミノ酸が並んでいると考えられる。したがって、二つの配列の遺伝学上の類似度を知るためには、上のように定義されたスコアの積算値が配列中の特定の領域で高くなっているかを調べる必要がある[3]。すなわち、配列中でスコアの積算値が最大値をとる領域を探し、その最大値をその配列の類似度と考える必要がある [10, 1]。配列ペアのスコア積算値が最大となる領域を Maximal Segment Pair (MSP) と呼び、そのスコアを次のように定義する。また、以下では、配列中の n 番目の文字ペアの類似度スコアを $S_n = s(X_n, Y_n)$ と表記する。

Definition 2 (MSP スコア). *2* つの長さ N の文字列が $\{X_n\}$、$\{Y_n\}$ があったとき、この文字列の *Maximal Segment Pair* のスコア S を次のように定義する。

$$S = \max_{1\leq i\leq j\leq N}\sum_{n=i}^{j}S_n = \max_{1\leq i\leq j\leq N}\sum_{n=i}^{j}s(X_n,Y_n) \tag{5}$$

[3]section 3.1 で見たように、実際のアラインメントの場合には、配列の中にギャップを挿入する必要がある場合があるが、ここではギャップを考慮しない。

どんな配列ペアにも MSP が存在するので、そのスコアの値が問題となる。Theorem 1 より二つの配列ペアが独立の場合には、$E[S] \leq 0$ となるので、二つ配列の MSP スコアが一定のしきい値 $s_0 > 0$ よりも大きくなった場合 $\{S \geq s_0\}$ には、二つの配列は遺伝的に深い関連があると判断する。しかし、得られたスコアが高い場合にも、本来は、まったく関係のない二つの配列が偶然スコアが高くなったり (false positive)、逆に、本来は関係の深い二つの配列が偶然スコアが低くなる (false negative) 可能性がある。したがって、これらの統計的誤差が十分小さくなるように、スコアのしきい値を設計する必要がある。

4 待ち行列理論と類似度スコアリング

4.1 Lindley equation

２つの配列 $\{X_n\}$ と $\{Y_n\}$ の類似度スコアを計算することを考える。待ち行列システムの待ち時間と対応させるために、W_n を次のように定義する。

Definition 3. 配列中の n 番目の文字を最後尾にした場合の *MSP* を W_n とする。すなわち

$$W_n = \max_{1 \leq i \leq n+1} \sum_{k=i}^{n} S_k = \max_{1 \leq i \leq n+1} \sum_{k=i}^{n} s(X_k, Y_k). \tag{6}$$

但し、$\sum$ の範囲が空集合の場合には、その値は 0 とする。

明らかに、$S = \max_{1 \leq n \leq N} W_n$ である。W_n は、待ち行列理論でよく知られた Loynes variable であり、Lindley の方程式を満たすことがよく知られている。

Theorem 2 (類似度スコアの Lindley の方程式). 文字列の類似度スコア W_n は次の方程式を満たす。

$$W_n = \max\{W_{n-1} + S_n, 0\}. \tag{7}$$

但し、$W_0 = 0$ とする。

Proof. 帰納法を使う。$n = 1$ の時には、

$$W_1 = \max_{1 \leq i \leq 2} \sum_{k=i}^{1} S_k = \max[0, S_1]$$

となり、成立する。次に $n-1$ の時に成立すると仮定する。

$$\begin{aligned}\max\{W_{n-1}+S_n,0\} &= \max\left\{\max_{1\le i\le n}\sum_{k=i}^{n-1}S_k+S_n,0\right\}\\ &= \max\left\{\max_{1\le i\le n}\sum_{k=i}^{n}S_k,0\right\}\\ &= \max_{1\le i\le n+1}\sum_{k=i}^{n}S_k\\ &= W_n\end{aligned}$$

□

先着処理順の単一サーバーのシステムへの客の到着間隔を T_n と客のサービス時間列を Z_n とすると、n 番目の客がサーバーでの処理を待つ時間 W_n は、次のように表される。

$$W_n = \max\{W_{n-1}+Z_n-T_n,0\}. \tag{8}$$

したがって、$S_n = Z_n - T_n$ と考えれば、(7) より、アミノ酸配列の類似度スコアの列と待ち時間は同一視することができる。

待ち行列理論では、$\rho = E[Z_n]/E[T_n] < 1$ のときシステムが安定し、待ち時間は定常分布を持ち、全稼働時間（サーバーが稼動する時間）が有限となることが知られている。対応するアミノ酸配列の類似度スコアの場合には、$E[S] < 0$ の場合には、全稼動機関に相当する MSP が有限となる。実際 Theorem 1 より独立なアミノ酸配列のペアでは、$E[S] \le 0$ となるので、遺伝子をコードしていないジャンク部分の配列が存在すれば、MSP は有限となる。

4.2 アミノ酸配列の W_n

Lindley equation (7) を使って、生命体のアミノ酸配列のスコア W_n を実際に計算した結果をいくつか示そう。

はじめに HBA HUMAN と呼ばれる遺伝子配列を別の 3 つの種類の遺伝子配列と比較する。図 1 では、この HBA HUMAN をスコア計算のベースになるアミノ酸配列として、アミノ酸置換マトリックス PAM250 でスコア計算を行った。HBB HUMAN は、HBA HUMAN と同種の遺伝子であることが知られている。HBB HUMAN ではスコア W_n がほぼ単調に大きくなっているのがわかる。対応する待ち行列システムで解釈すると、到着過程が、システムの処理能力を超え、システムが稼動し続けており、待ち時間が、徐々に増加・爆発していくことに対応している。遺伝子のスコアで考えれば、両方の遺伝子配列が完全に関連しており、ジャンク部分がなく、MSP が配列全体になっていることを示している。人工的に作られたランダムな配列 Random は、偶

然、似た配列が生まれることがあるが、ほとんどスコアの値が 0 になっていることもわかる。Random と比較すると、まったく HBA HUMAN とは異なる機能を持つ遺伝子配列である LUPLU は、比較的大きな値を持つ部分があるのがわかる。これは、LUPLU の配列が遺伝学上意味を持ち、ベースとなる遺伝子配列と比較的よく似た確率過程的な性質をもつためと考えられる。

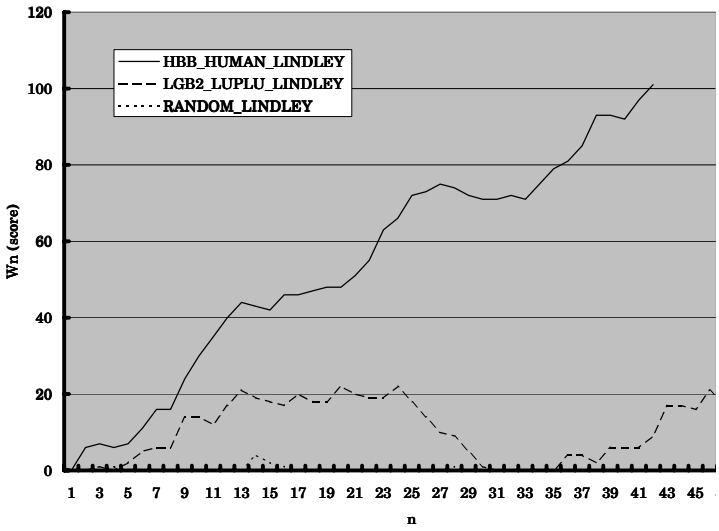

図 1: Sample path of Scores: The scores are calculated by PAM 250 based on the sequence of HBA HUMAN.

次の図 2 は、Section 3.1 でも紹介したレトロウィルスの遺伝子群の W_n である。比較ために、ランダムに入れ替えた配列 Random Arrange の W_n も示している。Section 3.1 で指摘したように、Unknown 以外のレトロウィルスの配列には、ジンクフィンガーと呼ばれる特徴的な部分があることが知られている。図 2 では Random Arrange 以外の配列は、ジンクフィンガーに相当する部分の W_n が高くなり、その特徴を捉えていることがわかる。より詳細に見ると、MMTV、RSV、SMRV の 3 つの配列は、ジンクフィンガー以外の部分もかなり W_n の値が高く、より「似た」配列であることがわかる。これに対して、Unknown と M MULV の二つは、ジンクフィンガーに挟まれている部分

の配列が「似ていない」ということもわかる。

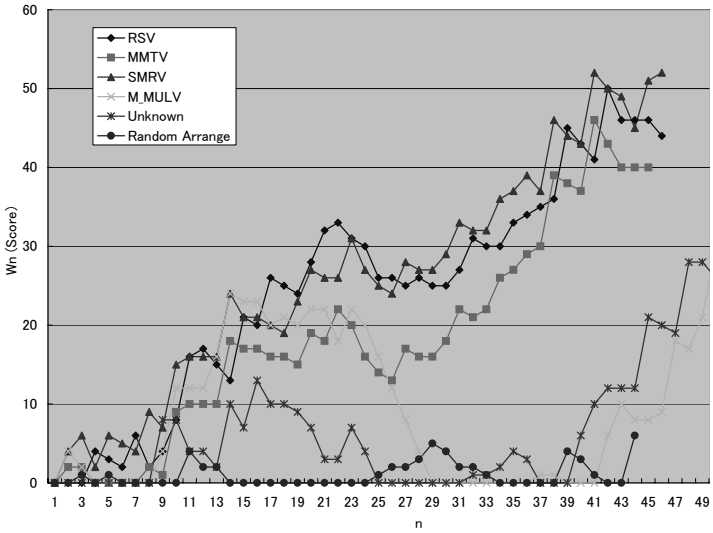

図 2: Sample path of Retrovirus: The scores are calculated by PAM 250 based on the sequence of HTLV. The sequences are Retorvirus' as shown in Section 3.1

これらの図をみてもわかるように、Lindley 方程式を使うことで、アミノ酸配列の遺伝的な特徴をとらえることが可能である。以下では、待ち行列モデルと対比することで、配列の類似度を定量的に判断する方法について検討する。

5　類似度スコア W_n のしきい値

5.1　待ち時間の Kingman Bound

どのような配列ペアにも類似度スコアが最大となる MSP が存在する。配列をランダムに入れ替え、まったく遺伝的な意味を持たないようにした配列であっても、W_n が偶然大きくなる可能性がある。したがって、類似度スコアがどの程度以上の場合に、アミノ酸配列ペアが似ていると判断するかが重要と

なる。はじめに、独立な二つのアミノ酸配列の類似度スコアがどのような分布をするかを考えよう。

アミノ酸配列の文字のスコアの列 $\{S_n\}$ が独立で同一な分布に従うとする。これは、遺伝的な情報はまったくコードされていない junk の部分に相当する。この場合、配列のスコアを表す W_n は、$GI/GI/1$ 待ち行列と考えることができる。$GI/GI/1$ 待ち行列の待ち時間分布については、Chernoff Bound を使って、その上限と下限 (Kingman Bound) が得られることが知られている（例えば、[12] P45 参照）。

Theorem 3 (Kingman Bound). $GI/GI/1$ 待ち行列システムの待ち時間 W_n は次の不等式を満たす。

$$\gamma e^{-\eta x} \leq P[W_n > x] \leq e^{-\eta x} \tag{9}$$

ここで、γ は 1 以下のある正の定数であり、η は W_n の *decay rate* と呼ばれ、次のようにして得られる。

$$\eta = sup\{s > 0 : E[e^{sS_n}] \leq 1\} \tag{10}$$

Remark 1. W_n の分布が $e^{-\eta x}$ の形で近似できることは、*BioInfomatics* でも知られている *([10] Theorem 1)*。

Corollary 1. 任意の $p \in (0,1)$ に対して、$w_0 = \frac{-\log p}{\eta}$ とすると

$$P[W_n > w_0] \leq p. \tag{11}$$

Proof. 仮定より、$p = e^{-\eta w_0}$ なので、Theorem 3 を使うと

$$P[W_n > w_0] \leq e^{-\eta w_0} = p.$$

□

Corollary 1 のように w_0 をとれば、ベースとなる配列と遺伝的に無関係なランダムな配列のスコア W_n がしきい値 w_0 を超える確率は p 以下となることがわかる。すなわち、配列スコアが w_0 よりも大きなスコアの場合に、その配列に遺伝的な意味があると判断すると、ランダムな配列が偶然大きなスコアをとり、その配列に遺伝的な意味があると誤って判断する確率を p に抑えることができる。

5.2 アミノ酸配列のしきい値

Section 3.1 のアミノ酸配列を例に、実際に W_n のしきい値を計算してみよう。図 2 でも使った塩基の順序をランダムに入れ替えて使った RandomArrange

配列のスコア S_n を使って $f(s) = E[e^{sS_n}]$ を次のように近似する。

$$f(s) \approx \frac{1}{N} \sum_{n=1}^{N} e^{sS_n}. \tag{12}$$

図 3 に、(12) のグラフを示す。この図から decay rate η は 0.3 となっていることがわかる。したがって、Corollary 1 より、スコア類似度のしきい値として $w_0 = 17.6611$ とすれば、$P[W_n > w_0] < 0.005$ となることがわかる。すなわち、HTLV とまったく無関係で遺伝的なつながりのない配列が w_0 を超える確率は $1/200$ であることがわかる。

図 3: $f(s) = E[e^{sS_n}]$:但し、S_n は RandomArrange とした場合

実際、図 2 を見ると、RandomArrange は、しきい値 $w_0 = 17.6611$ を超えていない。ところが、RSV、MMTV、SMRV の 3 つのレトロウィルスの配列は、$w_0 = 17.6611$ を超え、単調に増加しているのがわかる。これらの配列が偶然このように大きくなる確率はきわめて小さく、これらの 3 つのレトロウィルスの遺伝子は、非常に近い生物学的機能を実現していることが推定される。また、M MULV と Unknown について詳細にみると、8 から 25 くらいまでの配列の W_n のグラフの様子は極めて近い。しかし、Unknown はしきい値 $w_0 = 17.6611$ を超えていないので、遺伝的に意味のない配列が偶然レトロウィルスの機能を表す配列と似ていた可能性は否定できない。M MULV の場合、25 以降の配列で W_n が一旦かなり小さくなっている。この部分は、ジンクフィンガーの自由度の高いことを示している。RSV、MMTV、SMRV の 3 つと比較すると、M MULV のこの部分が確率過程的性質が異なるほどに大きな突然変異の影響を受けていると推定される。したがって、M MULV はこれらの 3 つと比べると遺伝的に遠い配列であることが予想される。また、40 以降は、M Mulv と Unknown は、ともにしきい値 $w_0 = 17.6611$ を超えている。したがって、この部分は、両者とも何らかの遺伝的機能を実現している可能性が高い。

6 Conclusion

本論文では、アミノ酸配列の遺伝配列の類似度スコアの確率過程的な性質を、待ち行列理論の Lindley 方程式を使って、待ち時間と対応させることができることを示した。また、GI/GI/1 待ち行列の待ち時間とランダムな配列を対応させ、遺伝学上意味のないアミノ酸配列が意味のある配列に偶然似る確率を導出できることを示した。

今後は、GI/GI/1 のような再生過程だけではなく、配列が自己相関を持ち、遺伝的機能を持つアミノ酸配列の待ち行列的なモデル化を詳細に行い、そこに待ち行列理論を応用できないかを検討する。

参考文献

[1] S.F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, Vol. 219, pp. 555–565, 1991.

[2] Andreas D. Baxevanis. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition*. John Wiley and Sons, 2001.

[3] Erhan Cinlar. *Introduction to stochastic processes*. Prentice-Hall, 1973.

[4] The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, Vol. 409, pp. 860–921, 2001.

[5] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. Atlas of protein sequence and structure. *Nat. Biomed. Res Found.*, 1978. Washinton DC.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 1998. ISBN: 0–521–62971–3.

[7] J. Craig Venter et al. The sequence of the human genome. *Science*, Vol. 291, pp. 1304–1351, 2001.

[8] David J. Galas. Making sense of the sequence. *Science*, Vol. 291, pp. 1257–1260, Feb 2001.

[9] Mikito Ishii and Hiroshi Kanehisa. Moji wo hikakushi naraberu. In *Human Genome Project and Knowledge Management*. Bifuukan, 1995.

[10] S. Karlin and S. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci USA*, Vol. 87, pp. 2264–2268, 1990.

[11] Leonard Kleinrock. *Queueing Systems*, Vol. 1. John Wiley and Sons, 1975.

[12] Leonard Kleinrock. *Queueing Systems*, Vol. 2. John Wiley and Sons, 1976.

[13] Shigeki Miyake and Hiroshi Kanehisa. *Human Genome Project and Knowledge Management*. Baifuukan, 1995.

[14] Sheldon M. Ross. *Stochastic Processes*. John Wiley and Sons, 1996.

[15] Ronald W. Wolff. *Stochastic modeling and the theory of queues*. Prentice-Hall, 1989.