DNA塩基配列の統計力学的尺度を用いた生物進化の解析

坂口 慶幸[1) 吉原 郁夫[2) 山森 一人[3) 剣持 直哉[4)

Criterion of Evolution Analyzing DNA Sequences based on Statistical Mechanics

Yoshiyuki SAKAGUCHI Ikuo YOSHIHARA Kunihito YAMAMORI
Naoya KENMOCHI

ABSTRACT

Eukaryotic genes are composed of exons and introns. The former are translated into protein, but the latter are not. All the functions of introns are not necessarily clear. However, introns are believed to have close relations to the evolution of species, therefore to analyze introns is of great importance. To reveal hidden properties in introns, the concept of energy is extended and the extended energy is estimated based on Boltzmann distribution that is equivalent to the conditional probability. To compare the extended energy with that of other species tells the extended energy increases as the elapse time of evolutionary branches. The experiments lead us to the conclusion that the extended energy is a promising criterion to distinguish if there exists of common characteristics of species that have a common ancestor.

Key Words:
Statistical Mechanics, Boltzmann Distribution, Intron, Ribosomal Protein Genes

1 はじめに

真核生物の遺伝子には，タンパク質合成に寄与するエクソンと呼ばれる部分とタンパク質合成に寄与しないイントロンという部分が存在するが，イントロンの方がエクソンよりもはるかに長い．タンパク質合成に寄与しないイントロンが，なぜ非常に長い塩基配列を持つのか，その存在理由や機能には未だ不明な点が多い．

原核生物の塩基配列はイントロンを持たないが，真核生物はイントロンを持っており，高等生物になるほど長くなっている．このことから，イントロンは進化の過程で別の部分が変化したものではないかと考えることができる．もしそうであるならば，イントロンにこそ進化の痕跡が残されている部分であると考えることができ，生物進化の解析にはイントロンを考慮に入れた解析を行うことが重要であると言える．このことから，本研究ではイントロンを含めた生物進化の解析を行う．

代表的な生物進化解析の手法として，塩基配列に対してアラインメントを施し，算出されたスコアを元に系統解析を行う方法がある[1)]．塩基配列から進化系統樹を作成するためのツールであるCLUSTALWは，この手法を用いる代表例である．

アラインメントとは，複数の配列を比較してある程度一致するパターンのことである．アラインメントを用いる手法は，生物進化の解析に有効な手法ではあるが，配列の一致を探すために全ての配列に対して全てのパターンの完全な探索を行わなければならず，このため計算量は膨大になり結果が出るまでに非常に長い時間がかかる．これを回避するため，複数の配列を比較する方法として，アラインメントを施す前に，配列全体である程度類似性の高い配列を取り出す手法を開発する必要性がある．

そこで，本研究では統計力学的な手法を用いて塩基配列全体が持つ擬似的なエネルギーを定義し，それを指標として生物種間で比較を行う手法を提案する．この手法が特にイントロンに対して有効であることを検

[1)]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)]フロンティア科学実験総合センター助教授

証するため，塩基配列全体でのエネルギー比較の他に，エクソンとイントロンを分け，それぞれでの生物種ごとに持つエネルギーの比較を行う．また，従来法との比較として，CLUSTALW を用いて作成した進化系統樹と比較を行うことによって，従来法のような配列の一致する部分を探す手法とは異なった，配列の全体での類似性を見るという本手法の有効性を検証する．

2 解析対象の概要

2.1 DNA の構造

DNA は，糖，リン酸，塩基から構成される．この中で糖やリン酸はどの生物種でも共通しているが，塩基は，アデニン（Adenine），グアニン（Guanine），シトシン（Cytosine），チミン（Thymine）の4種類から成っており，これらの並び方が生物種ごとに違う．使用される塩基配列は，それぞれの頭文字 A，G，C，T の4つのアルファベットの文字列として表すことで構成されている．

図1のように DNA は，転写領域と呼ばれるタンパク質の合成に関わる領域と，それ以外の非転写領域と呼ばれる生命現象に直接関与しない領域に分けることができる．

転写領域はさらに，エクソンと呼ばれる部分とイントロンと呼ばれる部分に分けられ，エクソンはスプライシングによってメッセンジャー RNA（mRNA）を構成し，それを翻訳することによってタンパク質が生成される．このことを遺伝子の発現と呼ぶ．一方，イントロンはスプライシングされる際取り除かれる部分であるため，タンパク質合成に直接寄与する部分ではない．しかし，先に述べたようにイントロンには何らかの進化の痕跡があるのではないかと考えられており，イントロンの解析を行うことは進化の解明をする上で重要であると考えられている．

2.2 リボソームとリボソームタンパク質

リボソームとは，図1のように mRNA を翻訳することによって，生命活動に必要なタンパク質を合成する部位で，地球上に現存するほとんどの生物に存在している．このリボソームは，3～4 種類の RNA と数十個のたんぱく質で構成される複雑な分子であるが，今までは RNA 部分（リボソーム RNA:rRNA）に多くの重要な機能が存在していると考えられていた．しかし最近の研究で，リボソームを構成するタンパク質部分（リボソームタンパク質）の異常に起因するとみられる疾患・変異が報告され始めている[2]．このことから，

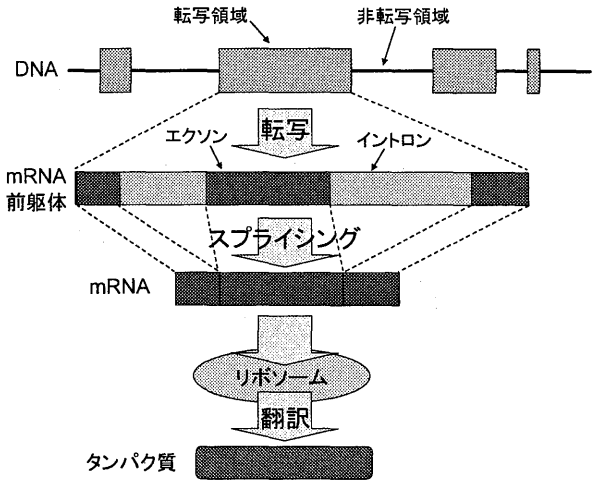

図. 1 遺伝子の発現

rRNA だけでなくリボソームタンパク質にも何らかの重要な機能が備わっているのではないかと考えられ，リボソームタンパク質遺伝子の解析も行われはじめている．本研究においてもこのような理由から，実験にはリボソームタンパク質遺伝子を使用する．

2.3 コドン

mRNA 塩基配列は，アミノ酸に翻訳され最終的にタンパク質が生成されるが，その際 mRNA は3塩基ずつ翻訳される．この3連続の塩基のことをコドンと呼ぶ．アミノ酸とコドンの対応は一対多であり，Grantham は遺伝子のコドン使用パターンに生物種による特徴があることを明らかにしている[3]．また，原核生物にはイントロンが存在していないということから，進化の過程でエクソンがイントロンに何らかの形で変化したのではないかと考えることができ，もしそうであれば，イントロンも3塩基1組で何らかの性質が見られるのではないかと考えられる．以上のことから，塩基配列の解析において単に1塩基ごとに解析するよりも，3塩基1組で考えた方が生物学的に意味が見出せるのではないかと考え，本研究では塩基配列を計算に用いる際，常に3連続塩基を1組にして考えることとする．

3 統計力学的尺度について

3.1 統計力学

統計力学とは，非常に多くの粒子の集合である物質を確率論と力学の原理を基礎にして物質全体の振る舞いを調べる学問である[4]．ここで，「粒子の集合である物質」を「塩基の集合である遺伝子」に置き換えることで，塩基配列全体の類似性を生物種間で比較できるだろうと考えられる．なぜなら，粒子と塩基どちらもたくさんの要素が相互作用をもたらしながら存在して

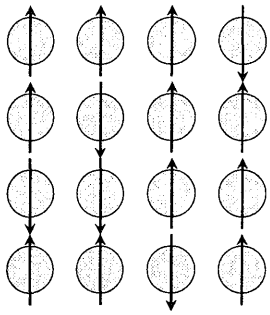

図. 2 イジング模型の例

いるという点で共通しているからである.

統計力学において相互作用のあるモデルとして代表的なものにイジング模型がある[5]. このモデルは図2のように, 各粒子が上向きか下向きのどちらかの磁気モーメントを持ち, このスピンが互いに相互作用することにより全体的には磁性を示す. このように, 磁気モーメントの状態は上向きと下向きの2状態であるので, それがそのまま情報処理の問題では0又は1に置き換えることが出来る. このことから, 統計力学は情報科学の分野と類似性があり, 情報科学の分野でよく使用されている.

3.2 本研究で用いる統計力学的手法

前節の方法は, 磁気モーメントの状態が2状態であるのでそれがそのまま0又は1に置き換えられるという簡単な方法であるが, 本研究の塩基配列が取りうる状態はA, G, C, Tの4つの状態であるため, 直接は利用できない.

そこで本研究では, 条件付確率と式 (1) で表されるボルツマン分布 f が等しい値をとることを利用する[6]. 各塩基が出現する条件付確率が計算可能なので, 各状態でのボルツマン分布が分かり, どの状態においても条件付確率とボルツマン分布の値が等しくなるような相互作用エネルギー E と温度が決定可能になる. ただしこの場合では条件付確率を用いているので, 状態は4状態から10状態に増える. ここで, 相互作用エネルギー E は, 相互作用を及ぼす2つの塩基 X と Y の間の距離が L であるときの相互作用エネルギーを $\varepsilon(X,Y)$ とすれば, 式 (3) より L が決まれば一意に決まるので, 距離の関数となって計算可能になる. しかし, 本研究での「距離」とは塩基間の実際の距離と対応するものではなく, 単純に何塩基離れているかという意味で用いる.

$$f = \frac{1}{1+e^{\beta E}} \tag{1}$$

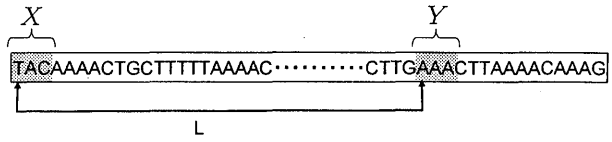

図. 3 塩基配列における条件付確率の計算

$$\beta = \frac{1}{kT} \tag{2}$$

$$E = \frac{\varepsilon(X,Y)}{L} \tag{3}$$

また, 式 (2) の β は, 温度に関するパラメータである. 本研究での温度は, 実際に計測した温度で, 式 (2) を計算するというものではなく, 相互作用エネルギーと同様1つのパラメータとして扱うものとする.

4 塩基間相互作用エネルギーの決定と解析のアルゴリズム

4.1 条件付確率の計算

はじめに塩基配列中に3連続塩基が出現する条件付確率を計算する. 図3において $X = \{AAA \sim TTT\}$ が出現したとき, その L 塩基先で $Y = \{AAA \sim TTT\}$ が出現する条件付確率を式 (4) に基づいて計算する. この L の長さを長くとるか短くとるかによって, 最終的にエネルギーを計算する範囲を決めることができる.

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{4}$$

4.2 塩基間相互作用エネルギー値の算出法

4.2.1 各3連続塩基に対して最適な塩基間相互作用エネルギー値の算出

X と Y の距離が L であるときの塩基間相互作用エネルギー $\varepsilon(X,Y)$ と仮想的な温度の値である β を式 (5) のように条件付確率とボルツマン分布が等しいことを利用して, 以下の手順で解く.

$$P(Y|X) = \frac{1}{1+e^{\beta E}} \tag{5}$$

Step1 X と Y の距離が L であるときの塩基間相互作用エネルギー塩基間相互作用エネルギー $\varepsilon(X,Y)$ と β の初期値を設定する.

Step2 2つの3連続塩基 ($X = \{X_1X_2X_3\}$ と $Y = \{Y_1Y_2Y_3\}$) の間に存在する塩基間相互作用エネルギー E を式 (6) に基づき計算する.

$$E = \frac{\sum_{a=1}^{3}\sum_{b=1}^{3}\varepsilon(X_a,Y_b)}{L} \tag{6}$$

Step3 ボルツマン分布 f の値を E を用いて計算し条件付確率 $P(Y|X)$ の値と比較する.

Step4 両者の差が小さくなるようにニュートン法を用いることによって $\varepsilon(X,Y)$ と β の値を変更する.

Step5 **Step1** から **Step4** を $P(Y|X)=0$ でないすべての3連続塩基の組み合わせについて行う.

4.2.2 塩基間相互作用エネルギー値の決定

前節で決定した初期値を用い，どんな3連続塩基に対しても式（5）を満たすような X と Y の距離が L であるときの塩基間相互作用エネルギー $\varepsilon(X,Y)$ と仮想温度 β の値を決める.

通常，最も簡単な方法として，単純にすべての3連続塩基に対しての $\varepsilon(X,Y)$ と β に対する平均を取ることが考えられるが，本研究の場合，求める $\varepsilon(X,Y)$（最終的には E）と β はボルツマン分布の式（1）の e の指数部にかかる値であり，求めるべき本来の値と少しでもずれてしまうとボルツマン分布の値は本来の値と大きくずれることになってしまう.

この影響を少しでも減らし，本来の値とのずれを可能な限り小さくするために単純な平均は取らず，遺伝的アルゴリズム（GA）を用いて値を決定する.

本研究で用いたGAの手順は図4に示す通りである.この図における初期個体生成とは **4.2.1** の初期値設定のことである.

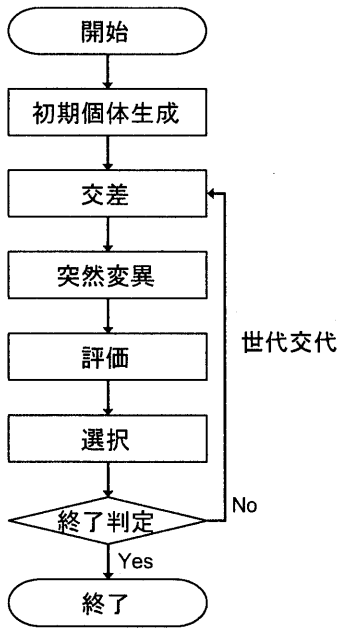

図. 4 GAの流れ

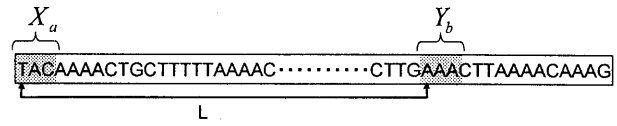

図. 5 塩基間相互作用エネルギーを用いた類似性比較

GAのパラメータ

- 初期個体:**4.2.1** で導出したすべての3連続塩基の組み合わせにおける X と Y の距離が L であるときの塩基間相互作用エネルギー $\varepsilon(X,Y)$ と仮想温度 β.
- 交差:交差率 100 %で一点交差.
- 突然変異:突然変異率 5 %で値をランダム増減.
- 評価関数:評価関数は以下の式（7）を用いる.

$$fitness = \frac{1}{|P(Y|X) - f| + 1} \tag{7}$$

- 選択:$fitness$ が高い順に個体を残す順位選択.
- 終了条件:$fitness$ = 1.0 になるかエリートに300世代以上変化が無かったとき.

4.3 配列全体が持つエネルギーの算出

ここまでで算出した X と Y の距離が L であるときの塩基間相互作用エネルギー $\varepsilon(X,Y)$ と仮想温度 β を用い，配列全体が持つ擬似的なエネルギーを比較する.このときエネルギーの比較は全てのデータの平均で行う.

図5において，2つの3連続塩基の距離を L 塩基取るときの塩基配列全体が持つエネルギー値 E_L を，**4.2.2** で決定した $\varepsilon(X,Y)$ と仮想温度 β に基づいて式（8），(9) のように計算し比較する．ここで G は塩基配列の長さである.

$$E_L = \sum_{i=0}^{G-L-2} E_i f = \sum_{i=0}^{G-L-2} E_i \frac{1}{1+e^{\beta E_i}} \tag{8}$$

$$E_i = \frac{\sum_{a=i}^{i+3} \sum_{b=i+L}^{i+L+3} \varepsilon(X_a, Y_b)}{L} \tag{9}$$

また，エクソンとイントロンを分けたエネルギー計算を行う場合には，ひとつの塩基配列からエクソンとイントロンをそれぞれ取り出し，それらを1つに繋げ，そのエクソンの部分配列，イントロンの部分配列それぞれについて上で述べた方法でエネルギーを計算する.

表. 1 使用する生物種

名称	略称
ヒト	Hs
マウス	Mm
ショウジョウバエ	Dm
センチュウ	Ce
出芽酵母	Sc
分裂酵母	Sp
マラリア原虫	Pf
メタン性古細菌	Mj
大腸菌	Ec
シロイヌナズナ	At

表. 2 使用するオーソロガス遺伝子

RPSA	RPS16	RPL11	RPLP1
RPS2	RPS18	RPL12	
RPS3	RPS20	RPL13A	
RPS5	RPS23	RPL17	
RPS9	RPL3	RPL23	
RPS11	RPL5	RPL23A	
RPS13	RPL7	RPL26	
RPS14	RPL8	RPL27A	
RPS15	RPL9	RPL35	
RPS15A	RPL10A	RPLP0	

5 統計力学的尺度を用いた類似性比較実験

5.1 実験に使用する遺伝子

実験に用いる遺伝子は，宮崎大学フロンティア科学実験総合センターリボソーム研究グループより提供されているリボソームタンパク質遺伝子のデータベース†に公表されているもの[7]を用いる．解析対象は，表1の10生物種である．これら各生物種のデータが完全に揃っているオーソロガス遺伝子表2の31種類を使用する．オーソロガス遺伝子とは，異なる生物種において相同な遺伝子座であり，構造上，機能上類似な遺伝子のことであり，生物種の進化に伴い共通の祖先遺伝子から進化した部分のことである[8]．

塩基間相互ポテンシャルエネルギーの算出には，ヒト（Hs）の遺伝子を用い，この値を用いて各生物種の塩基配列が持つ擬似的なエネルギーを計算することで，各生物種との比較はヒトを基準とした比較になる．つまり，ヒトとエネルギー値が離れれば離れるほど進化系統が離れていると考えることができる．

5.2 塩基間相互ポテンシャルエネルギー値計算結果

表3は，塩基間相互ポテンシャルエネルギー $E = \frac{\varepsilon(X,Y)}{L}$ と仮想的な温度の値である β を，全てのヒトのデータを $L = 100$ に対して算出しその平均値をとったものである．実際は，この他に $L = 20, 50, 200$ についても算出している．

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

表. 3 L=100 のときの平均値

	A	C	G	T
A	1.46	1.51	1.54	1.53
C		1.44	1.56	1.55
G			1.48	1.51
T				1.48
β	1.62			

5.3 実験結果・考察

5.3.1 結果

E_L を塩基長 G で割り，1塩基当たりが持つ平均エネルギーと定義して，それを生物種ごとにグラフにしたものが図6である．また，図7は塩基配列をエクソンとイントロンに分けた場合での1塩基当たりが持つエネルギーを生物種ごとにグラフにしたものである．

5.3.2 考察

図6において，高等生物であるほどエネルギー値も高くなっていることが分かる．ヒトを基準としてエネルギーを比較するためのパラメータを決定したため，ヒトと各生物種の持つエネルギーを比較すると，進化系統が離れているとされる生物種とは，エネルギーの差も大きくなっていることが分かる．

例外として，マラリア原虫（Pf）やシロイヌナズナ（At）の値が比較的高いエネルギー値を示しているが，この理由として，これらの生物には他の生物種に比べ繰り返しパターンが多く，このことが何らかの影響を及ぼしているのではないかと予想している．

また，L の値が大きくなるほど生物種間の値の差が大きくなってることが分かる．これは，L を短くとったときは，例えばイントロンに見られるGT-AGルールのような各生物種共通して存在する遺伝子自体のルールが反映されていることから，生物種ごとの違いが見られず，逆に，L を長くとると，生物種ごとの違いが現れているのではないかと考えられる．

図7からは，エクソンでのエネルギーの差はほとんどないことが分かる．このことは，はじめに述べた解析対象遺伝子であるリボソームタンパク質遺伝子の特徴と一致している．

図8は，従来手法の代表として，CLUSTALWを用いRPSA遺伝子のエクソンについての進化系統樹を作成したものである．図7のイントロンのグラフと図8の系統樹を比較すると，図7で一番値が近いヒト（Hs）とマウス（Mm）が図8でも隣同士であるように，進化系統樹で近い場所にいる生物種ほど，エネルギー値もある程度近いことが分かる．このことから，本手法

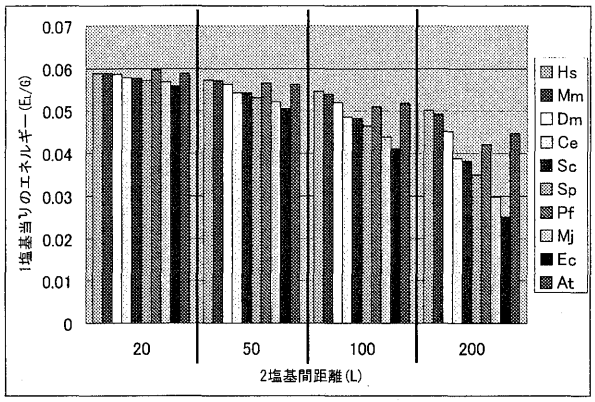

図. 6 各生物種が持つエネルギー平均値（1 塩基当たりで平均化）

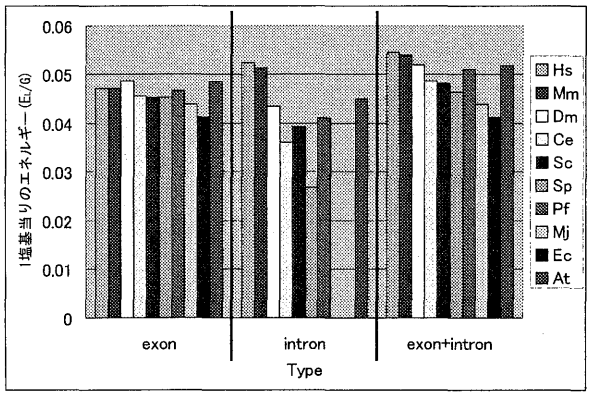

図. 7 エクソンとイントロンに分けた比較（L=100 の場合）

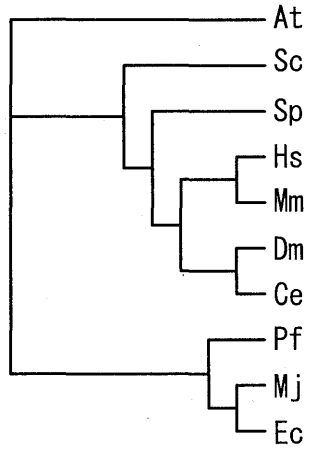

図. 8 CLUSTALW での RPSA エクソン部分の進化系統樹

はイントロンに対する生物進化の解析ができる手法であると言える．

6 おわりに

本研究では，統計力学的尺度として塩基配列に対し擬似的なエネルギーを定義し，それを指標として生物種間の比較を行う手法を提案した．エネルギー比較の際，本手法がイントロンのみでの生物種の比較が可能であるか検証するため，塩基配列全体での比較の他に，エクソンとイントロンを分け，それぞれでの生物種間におけるエネルギーの比較を行なった．また，従来法と全く異なった手法である本研究での手法が，生物進化の解析に有効な手法であるか確認するために，従来手法（CLUSTALW）で作成した進化系統樹との比較を行った．

リボソームタンパク質塩基配列を本手法に用い，その結果からリボソームタンパク質遺伝子のイントロンには，進化の順序関係が保存されていることが分かった．また，従来手法との比較によって，本手法がイントロンに含まれる進化の痕跡を発見する有効な手法であることが示された．

今後の課題として，今回実験により算出された結果に対して，塩基データの提供元であるリボソーム研究グループと協力し生物学的な検証を行うことが挙げられる．また，今回は統計力学的手法を用いるに当たり，塩基間の距離や温度などについて粗い当てはめを行っており，高い精度の実験結果を得られたとは考えにくい．よって実験結果の精度を上げるための改良を行っていくことが重要である．

参考文献

[1] 岡崎康司, 坊農秀雄: “バイオインフォマティクス 第2版”, メディカル・サイエンス・インターナショナル, 2005.

[2] 剣持直哉: “リボソームと疾患”, 実験医学, Vol. 22, No. 17, pp. 200–204 (2004).

[3] R. Grantham, C. Gautier, M. Gouy, R. Mercier and A. Pavé: “Codon catalog usage and the genome hypothesis”, Nucleic Acids Research, Vol. 8, No. 1, pp. r49–r62 (1980).

[4] 和田純夫: “熱・統計力学のききどころ”, 岩波書店, 1995.

[5] 久保亮五: “大学演習　熱学・統計力学”, 裳華房, 1961.

[6] 西森秀稔: “情報統計力学への招待”, 計測と制御, Vol. 42, No. 8, pp. 626–630 (2003).

[7] A. Nakao, M. Yoshihama and N. Kenmochi: “RPG: the Ribosomal Protein Gene database”, Nucleic Acids Research, Vol. 32, pp. D168–D170 (2004).

[8] 緒方宣邦, 野島博: “遺伝子工学キーワードブック 改訂第2版”, 羊土社, 2000.