自己組織化マップを用いた生物進化系統の解析

東 祐輔[1] 吉浜 麻生[2] 吉原 郁夫[3] 山森 一人[4] 剣持 直哉[5]

SOM-based Classification of Species using Sequences of Ribosomal Protein Gene

Yusuke HIGASHI Maki YOSHIHAMA Ikuo YOSHIHARA Kunihito YAMAMORI
Naoya KENMOCHI

ABSTRACT

The intron that does not contribute to protein synthesis has many unknown points for function and reason for existence, etc. It is thought that elucidation those points are useful for the clarification of the biological evolution and the genome function, etc. We tried analyzing intron by classification of species. The algorithm for classification used self-organizing map (SOM) that is able to view distribution of data in two-dimensional map. The using data for analysis is ribosomal protein gene that is thought useful for intron analysis. The codon frequency is calculated from each of transcriptional region and intron region, and classified species using SOM. In experimental result, it classified species in each of exon region and intron region. Therefore, it was understood that intron has unique information of species.

Key Words:
Self-Organizing Map, Intron, Ribosomal Protein Genes

1 はじめに

真核生物の遺伝子には、エクソンとイントロンの2つの領域が存在する。エクソンは、タンパク質合成に寄与するが、イントロンは、タンパク質合成に直接寄与していない。イントロンは、高等生物の遺伝子に一般的な配列で、通常、エクソンよりも多い。イントロンがどこからやってきて何のために広がったのか解明することは生物学上重要な課題である。

イントロンは、進化の過程で何らかの理由により挿入されたと考えられ、その中には生物進化の痕跡が含まれていると考えられる。イントロンに内在する生物進化の痕跡の発見には、各生物種のイントロンを比較することが有効であると考えられる。

従来ゲノム解析は、パターンマッチングのような、完全一致の手法がとられていた[1]。しかしながら、イントロンは、生物種ごとに配列長や塩基の並びが大きく異なっているため従来法の適応が難しい。

Granthamは、遺伝子のコドンの出現頻度に生物種による特徴が存在していることを多変量解析によって明らかにした[2]。また、阿部らは、コドンの使用頻度に基づいた自己組織化マップを作成し、生物種ごとにクラスタが生成されることを発見した[3]。

そこで、本研究では、イントロンに内在する進化の痕跡に関する手掛かりを得るため、自己組織化マップとコドンの出現頻度を使う。手掛かりを得る方法として、エクソン、イントロンの2つのコドンの出現頻度を算出し、それぞれの自己組織化マップを生成する。そして、生成された自己組織化マップをもとにイントロンに生物種に関する情報が含まれているかどうか調べる。また、エクソンとイントロンそれぞれのコドンの出現頻度から自己組織化マップを生成することで、エクソンに内在する特徴、イントロンに内在する特徴を調べる。本研究で、扱うデータは、リボソームタンパク質遺伝子の塩基配列である。

[1]工学研究科情報工学専攻学生
[2]フロンティア科学実験総合センター博士研究員
[3]情報システム工学科教授
[4]情報システム工学科助教授
[5]フロンティア科学実験総合センター助教授

2 遺伝子発現の機構とリボソームタンパク質遺伝子

2.1 DNAの構造

DNAは、ヌクレオチドと呼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。ヌクレオチドは4種類存在しており、塩基部分のみが異なっている。それぞれの塩基は、アデニン（A)、グアニン(G)、シトシン（C)、チミン（T）と呼ばれる。DNAは、糖とリン酸でできた主鎖に沿って塩基が並んでいるもので、2重螺旋構造を持っている。

DNAの中にタンパク質と機能性RNAの情報を含んでいる部分があり、遺伝子という。遺伝子には、エクソンとイントロンという領域が存在する。

2.2 遺伝子発現

遺伝情報（遺伝子）が読み取られてタンパク質ができることを遺伝子の発現という。遺伝子の発現のプロセスを図1に示す。まず、DNAから遺伝情報を含んでいる部分がRNA前駆体（pre-mRNA)に写しかえられる。これを転写という。次に、RNA前駆体からタンパク質に翻訳されない領域であるイントロン（Intron）が取り除かれ、メッセンジャーRNA（mRNA）となる。この操作をスプライシングという。リボソームがmRNAを読み込み、タンパク質を合成する。

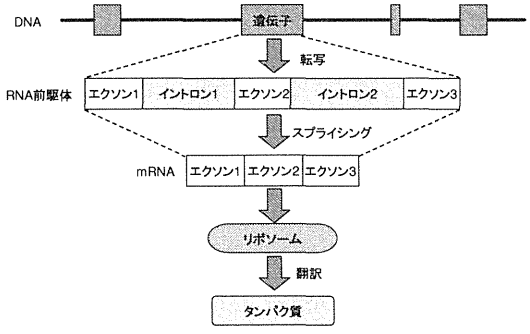

図. 1 遺伝子の発現

2.3 リボソームとリボソームタンパク質遺伝子

リボソームは、mRNAの翻訳を担う重要な細胞内装置で、それ自身がタンパク質とRNA（リボソームRNA）から構成された複合体で2つのサブユニットからなる。ヒトを含む高等動物において、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いたるまで大変よく保存されている[4]。

リボソームタンパク質遺伝子は、リボソームを構成するタンパク質である。すべての生物に存在し配列が保守的であることから、真核生物全体を通した生物進化系統の解析に適していると考えられている[5]。

なお、リボソームタンパク質遺伝子のデータは、RPG (Ribosomal Protein Gene database) に公開されている[6][7]。

3 自己組織化マップ

本研究で用いる自己組織化マップは、ニューラルネットワークの一種で教師なし学習アルゴリズムである。Kohonenにより記憶やその想起・連想のメカニズムを計算機上で実現するために開発された[8][9]。自己組織化マップは、高次元空間のデータを2次元空間に非線形写像するアルゴリズムである。そのため、高次元データの分布を2次元平面上に視覚化する有効なモデルである。

3.1 自己組織化マップの構造

自己組織化マップは、図2のように、入力層と競合層の2層構造である。入力層には、n個のノードがあり、それぞれが入力データの各要素に対応している。競合層のノードは、出力を視覚化するため、通常2次元に配列される。また、競合層のノードは重みベクトルを持っており、入力層のノード数に合わせn次元の要素を持っている。

3.2 主成分分析と一括学習自己組織化マップの併用

従来の自己組織化マップは、重みベクトルの値をある範囲でランダムに初期化しており、学習させるごとに生成されたマップが変わってしまう。また、データの入力順によっても出来上がるマップが変わってしまう。毎回生成されるマップが変わってしまうとマップ上での生物種の関係性を解析するのが困難になる。そのため、再現性のあるマップを生成する方法が必要となってくる。

本研究では、再現性のあるマップを得るため、阿部らが開発した主成分分析とデータの入力順に依存しない自己組織化マップ（一括学習自己組織化マップ）を併用する手法を導入する[10][3]。

3.2.1 主成分分析の利用

主成分分析とは、データの情報をなるべく落とさないようにデータの次元数を削減する方法で、線形写像を基礎としている。本研究で用いる初期化法では、第1主成分、第2主成分を用いる。

まず、n個の入力データを $\mathbf{x}^{(n)}(= x_1^{(n)}, x_2^{(n)}, \ldots, x_k^{(n)})$ と定義する。k は、入力データの次元数である。次

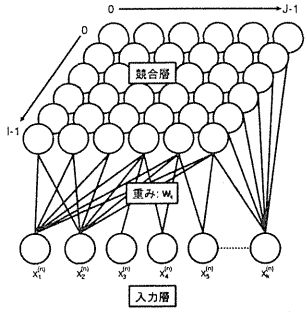

図. 2　自己組織化マップの構造

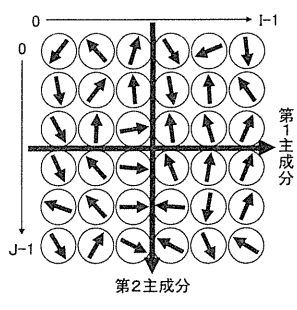

図. 3　主成分分析を用いた重みベクトルの初期化

に、競合層のノードが持つ重みベクトルを $\mathbf{w}_{ij}(= w_{ij1}, w_{ij2}, ..., w_{ijk})$ とする。このとき、競合層のノードの位置を $i(= 0, 1, .., I-1)$ と $j(= 0, 1, ..., J-1)$ で表す。重みベクトルの初期値を以下の式により定義する。

$$\mathbf{w}_{ij} = \mathbf{x}_{ave} + 5\sigma_1\left(\frac{i - I/2}{I}\right)\mathbf{b}_1 + 5\sigma_2\left(\frac{j - J/2}{J}\right)\mathbf{b}_2 \quad (1)$$

ここで、$\mathbf{x}_{ave}$ は入力データ全体の平均ベクトル、$\mathbf{b}_1, \mathbf{b}_2$ は第 1、第 2 主成分ベクトル、σ_1, σ_2 はこれら 2 つの軸に対する入力データ全体の標準偏差である。この式によって、図 3 のように、第 1 主成分と第 2 主成分を軸とした空間に 2 次元状の競合層を対応づけすることができる。この方法により、第 1 主成分、第 2 主成分を指標とした重みベクトルの値を決定することができる。ここで、図 3 のノードの中の矢印は、重みベクトルを表している。

3.2.2　一括学習アルゴリズム

一括学習アルゴリズムは、入力データの分類、重みベクトルの更新の 2 つのステップからなる。

Step1 入力データの分類

すべての入力データを最小のユークリッド距離を有する重みベクトルを持った競合層のノードに分類する。ユークリッド距離は、$E = \sqrt{\sum_k (x_k^{(n)} - w_k)^2}$ であるが計算上、次の式を使う。

$$E = \sum_k (x_k^{(n)} - w_k)^2 \quad (2)$$

6×6 の競合層に分類に 12 個の入力データを分類した図 4 を示す。格子は競合層、立方体は入力データを表している。

Step2 重みベクトルの更新

図 5 は、重みベクトルの更新の概要である。重みベクトルを更新する競合層のノードを中心とした範囲に含まれる分類された入力データから平均ベクトルを求める。次の式により重みベクトル $\mathbf{w}_{ij}$ を更新する。

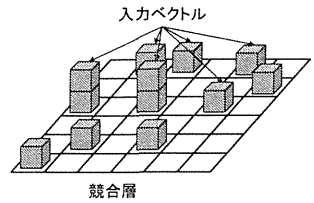

図. 4　入力データの分類

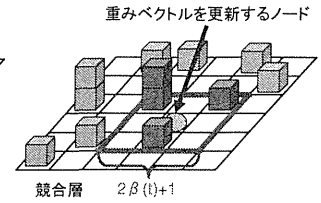

図. 5　重みベクトルの更新

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij}^{(old)} + \alpha(t)\left(\frac{1}{N_{ij}}\sum_{\mathbf{x}^{(k)} \in S_{ij}} \mathbf{x}^{(k)} - \mathbf{w}_{ij}^{(old)}\right) \quad (3)$$

S_{ij} は、i, j を中心とした一辺 $2\beta(t)+1$ の正方領域に分類された入力データ $\mathbf{x}^{(k)}$ の集合とし、N_{ij} は S_{ij} の要素数である。$t(= 1, 2, .., T)$ は学習回数を示す。また、$\alpha(t)$ は学習係数 $(0 < \alpha(t) < 1)$、$\beta(t)$ は近傍範囲を表している。$\alpha(t)$ および $\beta(t)$ を以下の様に定義する。

$$\alpha(t) = max\{0.01, \alpha_{init}(1 - t/T)\} \quad (4)$$

$$\beta(t) = max\{0, \beta_{init} - \beta_0 t\} \quad (5)$$

ここで、α_{init} と β_{init} は、それぞれの初期値とする。また、β_0 は、更新範囲の収束係数である。

4　ゲノムデータの扱い方

4.1　コドンを扱う理由

mRNA の塩基配列の連続した 3 個 1 組をコドンといい、それぞれのコドンが 1 つのアミノ酸に対応している。本研究では、ゲノムデータからコドンの出現頻度を算出し、それをもとに自己組織化マップを生成する。

4.2　コドンの出現頻度の算出

本研究では、コドンをカウントするときに、コドン同士の重複を許している。なぜならば、イントロンにおいて、コドン読みの開始位置がわからない。そのため、強制的にコドン読みした場合にイントロンに内在する情報を見落としてしまうと考えたためである。

コドンの出現頻度の算出の概要を図 6 に示す。図 6 は、イントロンのコドンの出現頻度を算出する場合である。まず、DNA 塩基配列の中にある遺伝子のイントロンを取り出し、それらをすべてつなげる。次に、つなげたイントロンから、1000 塩基づつ切り取っていく、このとき、それぞれの 1000 塩基は重複部分を持たない。次に、切り出した 1000 塩基からコドンをカウン

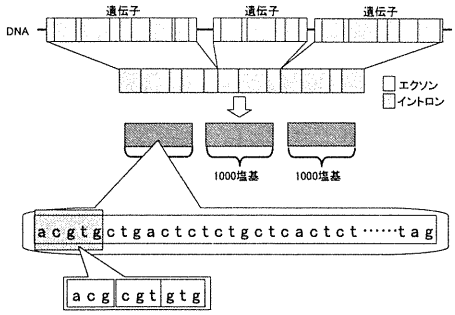

図. 6 コドンの出現頻度の算出の概要

トしていく、まず、先頭のコドンをカウントし、次に1塩基ずらしコドンをカウントする。例えば、acgtgの塩基配列からカウントされるコドンは、acg,cgt,gtgである。コドンの出現頻度は、A,G,C,Tの3連続塩基の出現頻度であるから、自己組織化マップへの入力データの次元は、$4^3 = 64$である。

5 イントロンに内在する特徴の解析

自己組織化マップによって、イントロンに生物種や進化に関する情報が含まれているか調べる。まず、エクソンによる自己組織化マップ、イントロンによる自己組織化マップをそれぞれ生成し、それぞれに、生物種に関する情報が含まれているかどうか調べる。次に、エクソンとイントロンに含まれる特徴を探すため、エクソンとイントロンそれぞれのコドンの出現頻度の自己組織化マップを生成する。

5.1 エクソンによる生物種の自己組織化マップ

5.1.1 実験条件

使用する生物種とそれぞれの入力データ（コドンの出現頻度）の数を表1に示す。コドンの出現頻度は、各生物種の塩基配列から切り出した1000塩基をもとに算出している。

表. 1 エクソンの各生物種の入力データ数

生物種名	ID	入力データ数
シロイヌナズナ	A	154
センチュウ	B	44
細胞性粘菌	C	38
ショウジョウバエ	D	72
ヒト	E	50
マラリア原虫	F	49
分裂酵母	G	69
出芽酵母	H	66
合計		596

また、自己組織化マップの各パラメータは次のようにした。

- 競合層のサイズ: 25×25
- 学習係数の初期値（α_{init}）: 0.5
- 更新範囲の初期値（β_{init}）: 16
- 学習回数: 100

5.1.2 実験結果と考察

エクソンによる自己組織化マップを図7に示す。マップ上の各アルファベットは、表1のIDに対応している。'-'はその点に何も生物種がプロットされなかったことを表しており、'*'は、複数の異なる生物種が同じ点にプロットされたことを表す。生成されたマップをみると、各生物種ごとにクラスタができていることがわかる。その原因として、エクソンは、タンパク質に寄与する部分であることから、各生物のコドンの出現頻度に特徴があるため、クラスタができたと考えられる。

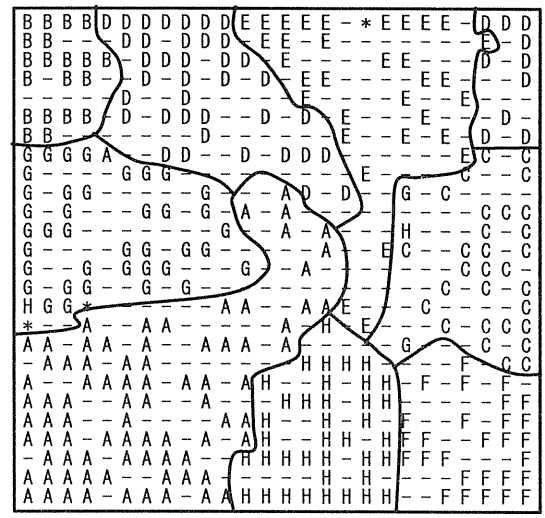

図. 7 エクソンによる生物種の自己組織化マップ

5.2 イントロンによる生物種の自己組織化マップ

5.2.1 実験条件

イントロンは、生物種間で塩基配列長の差が大きく、短いイントロンを持つ生物種が存在する。そのため、入力データの少ない生物種が存在しクラスタの確認が困難になると考えられる。そこで、入力データを増やし各生物種のクラスタを確認しやすくするため、コドンの出現頻度を算出する際にオーバーラップを許す方法をとった。詳しく述べると、コドンの出現頻度の算出の前に、1000塩基を取り出していたが、この取り出し位置をランダムし、入力データの数を増やした。使用する生物種とそれぞれの入力データ（コドンの出現頻度）の数を表2に示す。

表. 2　イントロンの各生物種の入力データ数

生物種名	ID	データ数
シロイヌナズナ	A	200
センチュウ	B	100
細胞性粘菌	C	100
ショウジョウバエ	D	100
ヒト	E	400
マラリア原虫	F	100
分裂酵母	G	100
出芽酵母	H	100
合計		1200

また、自己組織化マップの各パラメータは次のようにした。

- 競合層のサイズ: 34×34
- 学習係数の初期値（α_{init}）: 0.5
- 更新範囲の初期値（β_{init}）: 17
- 更新範囲の収束係数（β_0）: 1
- 学習回数: 100

5.2.2　実験結果と考察

イントロンの実験結果を図 8 に示す。生物種ごとにクラスタを生成していることがわかる。このことから、イントロンに生物種固有の情報が含まれていると考えることができる。

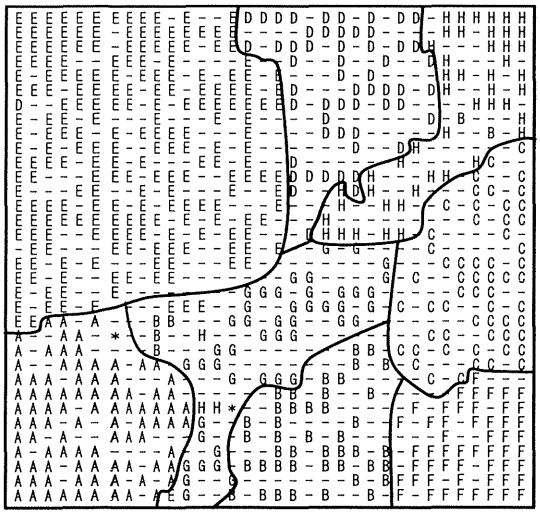

図. 8　イントロンによる生物種の自己組織化マップ

5.3　エクソンとイントロンによる生物種の自己組織化マップ

次に、エクソンに含まれる特徴とイントロンに含まれる特徴を調べるため、各生物種のエクソン、各生物種のイントロンのコドンの出現頻度による自己組織化マップを生成する。扱う生物種は 8 種類で、エクソンとイントロンに分けるので 16 種類のデータで自己組織化マップを生成することになる。

5.3.1　実験条件

各入力データは、コドンの出現頻度を算出する際に、オーバーラップを許す方法をとった。1000 塩基を切り出す際に、取り出し位置をランダムにする。各入力データ数を表 3 に示す。

表. 3　エクソンとイントロンの各生物種の入力データ数

生物種名	ID	入力データ数
シロイヌナズナ (イントロン)	A	200
センチュウ (イントロン)	B	100
細胞性粘菌 (イントロン)	C	100
ショウジョウバエ (イントロン)	D	100
ヒト (イントロン)	E	400
マラリア原虫 (イントロン)	F	100
分裂酵母 (イントロン)	G	100
出芽酵母 (イントロン)	H	100
合計		1200
シロイヌナズナ (エクソン)	I	200
センチュウ (エクソン)	J	100
細胞性粘菌 (エクソン)	K	100
ショウジョウバエ (エクソン)	L	100
ヒト (エクソン)	M	100
マラリア原虫 (エクソン)	N	100
分裂酵母 (エクソン)	O	100
出芽酵母 (エクソン)	P	100
合計		900

また、自己組織化マップの各パラメータは次のようにした。

- 競合層のサイズ: 45×45
- 学習係数の初期値（α_{init}）: 0.5
- 更新範囲の初期値（β_{init}）: 22
- 更新範囲の収束係数（β_0）: 1
- 学習回数: 100

5.3.2　実験結果と考察

実験結果を図 9 に示す。図 9 には、エクソンとイントロンの境界線を引いている。マップを見てみると、エクソンの領域とイントロンの領域ができているのがわかる。また、エクソン領域、イントロン領域共に、生物種に分かれているのがわかる。この結果から、大きく分けてエクソンとイントロンにはそれぞれ異なっ

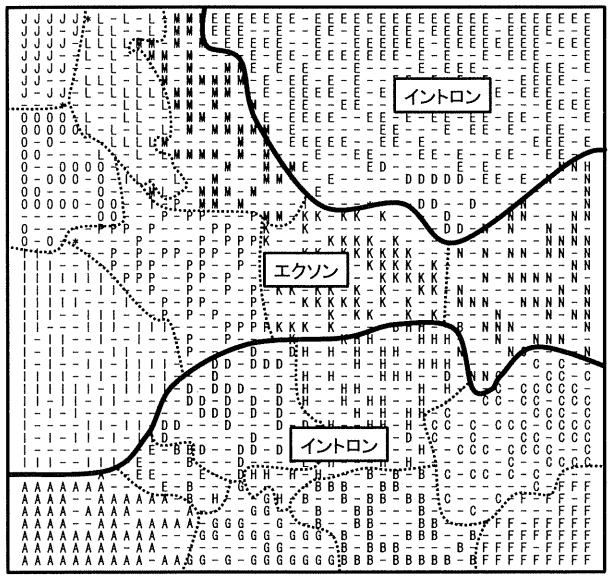

図. 9 エクソンとイントロンによる生物種の自己組織化マップ

た特徴が存在していること考えられ、さらに、それぞれの特徴の中には、生物種に関する特徴が内在していると考えられる。

次に、なぜイントロンとエクソンの領域ができたのか、その原因を探るためマップ上の各コドンの出現頻度の分布を可視化した。濃くなるほど、コドンの出現頻度が高いことを示し、薄くなるほど低いことを示す。

図 10 は、マップ上の AAG のコドンの出現頻度の分布を示しており、エクソンとイントロンの境界線を引いている。コドン AAG は、イントロンの領域では、出現頻度が低く、エクソンの領域では高いことがわかる。次に、コドン TTT の出現頻度の分布を図 11 に示す。図 10 とは逆に、イントロンの領域では出現頻度が低く、エクソンの領域では高いことがわかる。

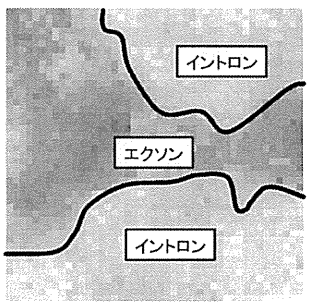

図. 10 コドン AAG の出現頻度の分布

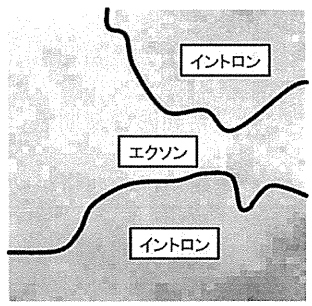

図. 11 コドン TTT の出現頻度の分布

6 おわりに

本研究では、イントロンに内在と思われる生物進化の痕跡に関する手掛かりを得るため、エクソンとイントロンそれぞれのコドンの出現頻度を算出し、自己組織化マップを生成した。得られたマップの生物種のクラスタを比較することでイントロンに含まれる特徴を調べた。

実験の結果、エクソン、イントロンのマップ共に、生物種のクラスタを生成した。この結果から、エクソン、イントロン共に生物種固有の情報が内在している可能性を示した。また、エクソンとイントロンの各生物種のデータからマップを生成した。その結果、マップにエクソンの領域、イントロンの領域ができ、それがコドン AAG、TTT に関係している可能性を示した。

今後の課題として、4 連続塩基や 5 連続塩基の出現頻度からマップ生成することや、多くの生物種を扱うことが挙げられる。

謝辞

本研究の一部は、文科省科研費・若手(B)No.17770207 及び基盤 (C)No.17500146 により行われた。

参考文献

[1] Cyntbia Gibas, Per Jambeck, 実践バイオインフォマティクス-ゲノム研究のためのコンピュータスキル-, 株式会社オライリー・ジャパン, 2002

[2] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res, 8:r49-462

[3] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura,Informatics for unveiling hidden genome signatures,Genome Research, 13:693-702

[4] 剣持 直哉, リボソームと疾患, 実験医学 増刊, 株式会社羊土社, 22: 200-205. 2004

[5] 五条堀 孝 編, ゲノムからみた生物進化の多様性と進化, シュプリンガー・フェアラーク東京株式会社, 2003

[6] http://ribosome.med.miyazaki-u.ac.jp

[7] Akihiro Nakao, Maki Yoshihama, Naoya Kenmochi, "RPG: the Ribosomal Protein Gene database", Nucleic Acids Res, 32(Database issue):D168-170. 2004

[8] T. コホネン, 自己組織化マップ, シュプリンガー・フェアラーク東京株式会社, 1996

[9] 徳高平蔵, 岸田悟, 藤村喜久郎, 自己組織化マップの応用, 海文堂, 1999

[10] T. Abe, S. Kanaya, M. Kinouchi, Y. Kudo, H. Mori, H. Matsuda, C.D. Carpio, T. Ikemura, and T. Ikemura, Gene classification method based on batch-learning SOM, Genome Informatics, 10:314-315, 1999