統計数理（2004）
第 52 巻 第 1 号 207–215

公開講演会要旨

ゲノム DNA 配列に潜んでいる生物種の個性を明らかにする新規な統計数理的手法

阿部 貴志[1,2,3] ・金谷 重彦[4] ・木ノ内 誠[5] ・池村 淑道[1,6]

（2003 年 11 月 5 日，統計数理研究所 講堂）

要　　旨

21 世紀に入り，ゲノム塩基配列の解読は益々加速する勢いにあり，データベースに既に収録されている配列についてさえ，個々の研究者を圧倒するほど大量に見える．その大量情報の全体から見えてくる未知の基本的な知識を得ることが，ゲノム科学における重要な課題である．我々は，ヘルシンキ大学のコホネンにより，記憶やその想起・連想のメカニズムを計算機上で表現するために開発された自己組織化マップ法（SOM）に着目した．教師なしのニューラルネットワークアルゴリズムであり，大量で複雑な情報について，似た情報を自ずと集める（自己組織化する）ことを計算機上で実現する．ゲノム塩基配列の解読の進んだ 13 種類の真核生物のゲノム配列を対象に，10 kb 及び 100 kb の断片配列内の 3 連と 4 連塩基の出現頻度について，SOM 解析をおこなった．生物種に関する情報を計算機に与えていないのに，大半の断片配列が生物種ごとに分離しており，各生物種のゲノム配列に潜む連続塩基頻度の特徴を的確に検出していた．ゲノムに存在する機能的に重要なシグナル配列の計算機探索が可能になること，自然環境に存在する培養が困難な微生物類の生物多様性や系統を明らかにする新規な視点でのゲノム解析が可能になること等を紹介した．

キーワード：ゲノム塩基配列，バイオインフォマティクス，自己組織化マップ（SOM），連文字頻度解析，教師なしニューラルネットワークアルゴリズム．

1. はじめに

生命の設計図であるゲノムは，どの生物種でも 4 種類の塩基（A, T, G, C）で書かれている．例えばヒトのゲノムの場合は，30 億塩基（3 ギガ塩基；3Gb）で書かれており，23 本の染色体から構成されるので，23 の長文からなると表現できる．日本を含む世界各国において，ゲノム塩基配列の解読は益々加速する勢いにあり，200 を超える生物種のゲノムの完全配列が解読され，

[1] 国立遺伝学研究所 集団遺伝研究系：〒411–8540 静岡県三島市谷田 1111
[2] 総合研究大学院大学 遺伝学専攻：〒411–8540 静岡県三島市谷田 1111
[3] ザナジェン：〒213–0012 神奈川県川崎市高津区坂戸 3–2–1 KSP D-1037
[4] 奈良先端科学技術大学院大学 情報科学研究科：〒630–0101 奈良県生駒市高山町 8916–5
[5] 山形大学 応用生命システム工学科：〒992–8510 山形県米沢市城南 4–3–16
[6] 現 総合研究大学院大学 葉山高等研究センター：〒240–0193 神奈川県三浦郡葉山町（湘南国際村）

総計で約 34.3 Gb の塩基配列がデータベースに収録されている．この大量なゲノム情報から，多様で多量な知識を効率的に発掘することが，ゲノム研究の重要な課題である．大量で複雑な情報からの知識発見には，統計数理的な解析が有効な手段となる．我々のグループは，塩基配列を文章として扱い，単語の出現頻度解析(Word Count)を行うことで，ゲノム配列に潜む多様な情報を効率的に抽出可能なことを見い出した(Abe et al. (2003))．ここで単語とは，2 連・3 連・4 連塩基のような連続塩基を意味する．2 連塩基については，AA, AC, AG, . . . , TT であり，4×4 の 16 種類の単語として扱える．例えば，以下のような短い文章(配列)が与えられた場合，

AACGACAACGCCGG

2 連塩基では，AA が 2 回，AC が 3 回，CG が 3 回，GA が 1 回 , . . . , 3 連塩基では，AAC が 2 回，ACG が 2 回，CGA が 1 回，GAC が 1 回 , . . . のように，一文字ずつずらしながら連続塩基(単語)を数える．

2 連塩基の頻度は 16 次元，3 連塩基では 64 次元，4 連塩基では 256 次元のベクトルとして表現できる．100,000 個の文章(配列)が与えられた場合には，これらは多次元空間の 100,000 個のデータである．Word Count は統計数理の分野では馴染みのある課題であるが，このような単純な解析を，データベースに収録されているゲノム配列の全体を対象とし大規模に行うことで，思いがけない新規な知識発見が可能になった．我々のグループが行っているゲノム情報学の研究を中心に，この分野の現状を紹介する．

2. 自己組織化マップ(Self-Organizing Map; SOM)のゲノム情報解析のための改良

我々は，上記の多次元空間の大量な単語頻度データを解析する上で，コホネン(Kohonen (1982, 1990), コホネン(1993), Kohonen et al. (1996))が記憶やその想起の機構を研究するために開発した，教師なしニューラルネットワークアルゴリズムである自己組織化マップ(SOM)に着目した．記憶やその想起は，複雑な情報の処理を効率的に行っている典型例であり，大量なゲノム情報からの知識発見の方法を考える上で重要なヒントを与える．SOM は大量で複雑な情報について，似た情報を自ずと集める(自己組織化する)ことを計算機上で実現している．記憶の研究を目的に開発されたが，応用性の高い情報学的手法として知られるようになった．工学・経済学・言語学のような大量で複雑な情報を解析する分野で普及してきたが，塩基配列の解析には殆ど用いられずにきた．比較的長い計算時間を必要とし，出来上がった地図がデータの入力順に依存する問題があった．記憶の場合には，情報の入力順序に依存することに意味があるが，ゲノム配列の解析では，対象とするゲノムの種類が同じならば，どのグループが解析しても，再現性の高い地図を得ることが重要になる．我々は,「学習過程と得られる地図がデータの入力順序に依存しないようにする」という特徴を SOM に導入した(Kanaya et al. (2001), Abe et al. (2002, 2003), 金谷 他(2003), 池村(2002))．

全ての生物種のゲノムは同じ 4 種類の文字で書かれているので，断片的な配列(例えば 10,000 塩基, 10kb)が多数与えられただけでは，何種類の生物種のゲノムに由来するのか，どの生物の配列なのかを知ることは不可能に思える．しかしながら，上述の連続塩基(単語)の出現頻度は生物種間で明瞭に異なっており，SOM を用いることで，断片配列を生物種ごとに高精度に分離することが可能であった(Abe et al. (2002, 2003))．データの入力順序に依存しない一括学習型 SOM (Batch-learning SOM)を用いて，3 連と 4 連塩基頻度に着目し，ゲノム配列の解読の進んでいる 13 種類の動植物と真核微生物のゲノム配列(総計で約 3 Gb)を解析した例を，図 1A に紹介する．各々の生物種のゲノム配列を 10 kb ごとに断片化して得た合計で約 300,000 断片配列(文章)，ならびに 10 kb ごとに移動させた重複を含む 100 kb の 300,000 断片配列の全体

について，3 連塩基($4^3 = 64$ 種類の単語)または 4 連塩基($4^4 = 256$ 種類の単語)の出現頻度を算出し，SOM 解析を行った．64 ならびに 256 次元空間の 300,000 データについての SOM であり，生物種の情報を与えなくても，生物種ごとに断片配列が自己組織化し，分離していた．具体的な計算過程を以下 3 節で方法として紹介するが，まず結果を知りたい読者は，この部分を飛ばして，4 節の「ゲノム配列に潜む生物種に特徴的なサイン(Genome Signature)の実体」へ進んでほしい．

3. 方法

SOM は，初期値設定・学習・分類の三つのステージからなる．個々の断片配列(例えば 10 kb)における連続塩基頻度(この場合は，3 と 4 連塩基頻度)によりベクトルを定義する．第 k 番目の断片における頻度組成を $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kM})$ とするが，M はベクトルの次元数であり 3 連塩基の場合は 64 である．ここで $k = 1, 2, \ldots, N$ であり，N は入力断片配列の総数で，この場合は 300,000 となる．2 次元の格子点を i, j $(i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J)$ とする．この例では，I を 200 に設定している．J については後に述べるが I に依存して設定する．2 次元の格子点 (ij) 上に，M 次元のリファレンスベクトル $\boldsymbol{w}_{ij}$ の初期値を，以下の式により定義する．

$$\boldsymbol{w}_{ij} = \boldsymbol{x}_{\mathrm{ave}} + 5\sigma_1 \boldsymbol{b}_1 \left(\frac{i - I/2}{I} \right) + 5\sigma_2 \boldsymbol{b}_2 \left(\frac{j - J/2}{J} \right) \tag{3.1}$$

$\boldsymbol{x}_{\mathrm{ave}}$ は入力ベクトル $\boldsymbol{x}_k$ $(k = 1, 2, \ldots, N)$ の全体の平均値ベクトル，b_1 および b_2 はそれぞれ主成分分析(PCA)の第 1 および第 2 主成分ベクトルである．記憶の問題を主目標にする従来型の SOM の場合，初期ベクトルはランダム値として定義するのが通常であるが，ゲノム配列の解析においては再現性が重要になる．この観点から，我々は，主成分分析の第 1 と第 2 主成分の軸を基礎に，(3.1)式で初期ベクトルを設定している．主成分分析は，比較的少数の生物種の遺伝子塩基配列の解析に有用であり，生物種の特徴を反映した分類が得られることから(Kanaya et al. (1996, 1999))，初期ベクトルの作成に使用した．σ_1 および σ_2 はこれら 2 軸に対する入力ベクトル全体の標準偏差であり，$J = I\frac{\sigma_2}{\sigma_1}$ として，初期ベクトルを設定する．

学習としては，始めに全ての入力ベクトル $\boldsymbol{x}_k$ を最小のユークリッド距離を有するリファレンスベクトル $\boldsymbol{W}_{i'j'}$ に分類する．つぎに(3.2)式によりリファレンスベクトル $\boldsymbol{W}_{ij}$ を更新する．

$$\boldsymbol{W}_{ij}^{(new)} = \boldsymbol{W}_{ij}^{(old)} + \alpha(t) \left(\frac{\sum\limits_{\boldsymbol{x}_k \in S_{ij}} \boldsymbol{x}_k}{N_{ij}} - \boldsymbol{W}_{ij}^{(old)} \right) \tag{3.2}$$

ここで，近傍 S_{ij} は $i - \beta(t) \le i' \le i + \beta(t)$ かつ $j - \beta(t) \le j' \le j + \beta(t)$ の条件を満たす格子点 $i'j'$ に分類された入力ベクトル $\boldsymbol{x}_k$ の集合，N_{ij} は S_{ij} の要素数，t $(= 1, 2, \ldots, T)$ は第 t 回目の学習回数を示す．また，$\alpha(t)$ は学習係数 $(0 < \alpha(t) < 1)$，$\beta(t)$ は近傍を決定する数である．SOM で一般に使用されている以下の $\alpha(t)$ および $\beta(t)$ を採用している．

$$\alpha(t) = \max\left\{ 0.01, \alpha_{init} \left(1 - \frac{t}{T} \right) \right\} \tag{3.3}$$

$$\beta(t) = \max\left\{ 1, \beta_{init} - t \right\} \tag{3.4}$$

α_{init} は，学習ごとに設定する学習係数 $\alpha(t)$ の初期値とする(例えば $\alpha_{init} = 0.9$)．T は総学習回数を示す．ここで，β_{init} は，学習ごとに設定する近傍領域数 $\beta(t)$ の初期値とする(例えば $\beta_{init} = I/4$)．

ここまでを，1 回分の学習とし，設定した学習回数(例えば $T = 100$)だけ繰り返し学習を行

う．学習は次式によって定義する二乗誤差で評価する．

$$e(t) = \sum_{k=1}^{N} \{\boldsymbol{x}_k - \boldsymbol{w}_{i'j'}\}^2 \tag{3.5}$$

ここで $\boldsymbol{w}_{i'j'}$ は $\boldsymbol{x}_k$ と最近隣の位置にあるリファレンスベクトルである．改良型 SOM プログラムは，Xanagen 社（神奈川県川崎市高津区坂戸, http://www.xanagen.com）により提供されている．

4. ゲノム配列に潜む連文字頻度の生物種による特徴的なサイン（Genome Signature）

図 1（A）に示すように，計算機に生物種の情報を与えていないのに，連続塩基の頻度パターンの類似度のみで，断片配列の大半が生物種ごとに高精度に分類されていた．単一の生物種のみの配列からなる格子点は生物種ごとに色分けし，異なる生物種に由来する配列が混在する格子点は黒とした．ヒト（図 1（B）の H），フグ（F），ゼブラフィシュ（Z），イネ（R），シロイヌナズナ（A），線虫（C）はゲノムサイズが大きいので，多数の断片配列を与えており，SOM 上で大きな領域を占有する．異なる生物種の配列が混在する格子点（黒）は，100 kb の SOM では殆ど見られない．10 kb の配列より，100 kb の配列の方が分離能が高いことを示す．例えば，ヒト配列の場合，10 kb 配列の SOM では 97%以上が，100 kb の配列では 99%以上が，ヒトの領域（H）内に位置していた（この解析では，ヒトゲノムの場合，配列の解読がほぼ完了した 10 本の染色体の配列のみを用いている）．

計算機が行った分類が種による分類と合致していたことは，大半の断片配列の内部には，各生物種を特徴づける単語の使用頻度に関する個性が存在し，計算機がその生物のサイン（signature）のように識別したことを意味する．文章の Word Count の例としてこの内容を説明しよう．例えば，20 人の米国の大統領について，各々500 件で，合計 10,000 件の演説原稿を集めたとして，200 種類の単語（例えば，power, military, peace, economy, university, science, women, music, ...）に着目した SOM を考える．各大統領には言葉の好みが存在するはずであり，各原稿について大統領の名前を計算機に教えていなくても，単語の使用頻度分布の似た原稿を近接させる（200 次元空間での距離の近いデータを地図上で近付ける）ことで，各大統領の原稿群がクラスタ化すると考えられる．この演説原稿の例については，どのような種類の単語に着目するかで，大統領間の分離能は変わってくるはずである．一方，ゲノムの連続塩基の頻度解析では，3 連か 4 連かの選択の余地はあるが，それが設定されれば，後は全ての連続塩基（単語）が解析の対象になり（3 連塩基では 64 種類の単語），曖昧さのない解析が可能となる．図 1（A）の例では，4 連塩基の方が，3 連塩基よりもやや分離能が良い．連塩基の長さの選択は，解析する配列の長さと関係する．我々は，5 連塩基の 1024 次元空間や 6 連塩基の 4096 次元空間を解析したことがあるが，多次元になるに従い計算時間が膨大になるので，別種の工夫が必要と考えられる．

5. ゲノムサイン（Genome Signature）の実体とその生物学的な意味

大統領の演説原稿を対象にした SOM を行った場合，計算機は各大統領の言葉の好みを検出し，クラスタリングを行うと考えられる．それでは，各生物のゲノムの単語の好みはどのようなものであろうか．計算機がどの連塩基に注目したのかを知る方法を，図 1（B）で紹介する．各連文字について，SOM のどの領域の格子点群で好かれ，どの領域で嫌われていたのかを数値化し，赤色と青色で表示した．ランダムな配列から予想される頻度に近い場合は，白色で示す．図 1（B）で紹介する連塩基において好みの程度が変化する部位は，図 1（A）の生物種の境界に一致していた．重要なことは，単一の単語よりは，複数の単語の組み合わせパターンでゲノ

ム個性がより鮮明になる点にある．それがゲノムサイン(Genome Signature)の実体をなすが，その生物学的な機能や，形成された進化機構が興味深い．生物種固有の特徴を生む原因としては，以下のようなものが考えられる．

1) 突然変異やその修復機構と関係する．G+C%として表現されるゲノムの塩基組成は，変異や修復機構を反映する．
2) 危険性のある(有害な)配列を避ける．原核生物の場合，その生物種の持つ制限酵素が切断する 4 塩基が特徴的に低頻度で，クラスタ化の要因であった(Abe et al. (2003))．メチル化酵素を持つとはいえ，危険な切断配列が特徴的に低頻度である．
3) 広範囲の生物種，特に高等真核生物においては，反復配列がゲノム上に散在しており，総計としてゲノムの大きい領域を占めることがある．但し，生物種別にクラスタ化する上で，反復配列が主要因でないことが判明している．
4) 重要なシグナル(例えば，転写因子と配列特異的に安定な結合をするシグナル)は，塩基組成から得られるランダム配列の予測値から明瞭にずれる．転写因子に対して高い配列依存性を示し，安定に結合するシグナル配列は，ランダム配列からの予想値よりも低頻度に出現する傾向が見られた．

統計数理ならびに情報学的視点から，暗号文を含む文書類の単語出現頻度解析(Word Count)として，4)は特に興味深い．ATGC からなる長文(ゲノム配列)だけが与えられても，各生物が重要にしているシグナルを推定できる可能性が示唆されている．これからのゲノム科学は，ポストゲノムと呼ばれる時代を迎えつつある．塩基配列の解読が行われるが，通常の分子生物学や生化学の実験が殆ど行われないゲノムが急増している．従来は実験的な研究が行われてきた課題を，極力 *in silico* 実験(計算機を用いた解析)で代行することが重要になる．教師なしのアルゴリズムである SOM は，正にこの目的に合致している．ゲノムに存在する機能上重要なシグナル類は，通常は 3 連や 4 連塩基よりは長い場合が多い．広範囲のゲノムについて，5-8 連塩基へと SOM 解析を進めれば，シグナルの候補配列の *in silico* 探索が可能と考えられる．

6. 多様な環境に生息する生物種の多様性と生物系統に関する新規な研究法

SOM を用いた実用的な研究の例を以下に紹介する．多様な地球環境に生息する多種類の生物種を対象にした，新しい視点でのゲノム解析技術が発展している．極限環境を含む多様な環境で生育する微生物類については，培養が困難な例が大半を占めている．例えば，深海の海底に存在する火山口の近くでは，150 度近くの高温で生育する微生物が知られており，生命の起源との関係でも注目を集めている．これらは，実験室で培養が行えない難培養性微生物の典型例である．また，南極の古い時代の氷の層内に閉じ込められた微生物の存在が知られており，我々の腸内においてすら，多様な難培養性微生物の存在が確認されている．実験室で培養が行えない難培養性微生物類については，通常の実験的な研究が困難であり，殆ど研究されずにきた．言い換えれば，新規で興味深い遺伝子類を豊富に保有する可能性があり，産業的にも注目を集めている．難培養性微生物を解析する新しい技術として，環境中で生育する生物群を含有する試料から培養を行わずに直接的にゲノム DNA の混合物を抽出し，配列決定を行う技術が開発されている．しかしながら，得られた多数の断片配列の集合のみでは，その試料に存在する生物の種類，系統群，それらの新規性等を推定することは困難であった．

連続塩基頻度のみに着目することで，断片配列を生物種ごとに高精度に分類可能な SOM の特徴に基づき，環境微生物ゲノムの多様性や新規性を推定するための新しい系統分類法，新規性の高い未知微生物ゲノムを効率的に探索する手法の開発が可能である．図 2 の例では，約 65

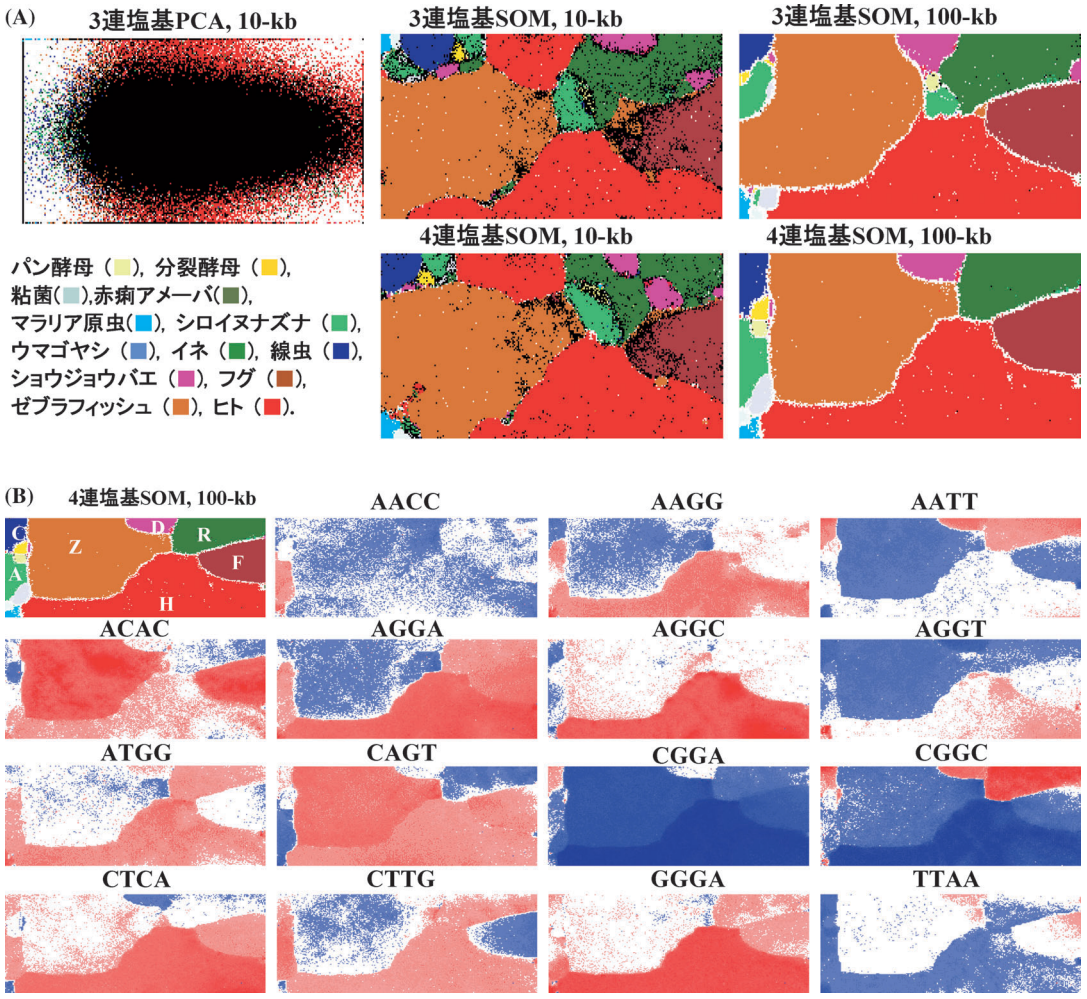
(A)
3連塩基PCA, 10-kb
3連塩基SOM, 10-kb
3連塩基SOM, 100-kb
4連塩基SOM, 10-kb
4連塩基SOM, 100-kb
パン酵母（ ）, 分裂酵母（ ）,
粘菌( ),赤痢アメーバ( ),
マラリア原虫( ), シロイヌナズナ（ ）,
ウマゴヤシ（ ）, イネ（ ）, 線虫（ ）,
ショウジョウバエ（ ）, フグ（ ）,
ゼブラフィッシュ（ ）, ヒト（ ）.
(B)
4連塩基SOM, 100-kb
C
D
R
Z
F
A
H
AACC
AAGG
AATT
ACAC
AGGA
AGGC
AGGT
ATGG
CAGT
CGGA
CGGC
CTCA
CTTG
GGGA
TTAA

図 1. **（A）真核生物 13 種の 10 kb ならびに 100 kb 配列について，3 連および 4 連塩基の出現頻度について，SOM を行った際の生物種ごとの分布図．色付きの格子点は 1 つの生物種だけからなるものであり，生物種と色との対応は図中に示した．複数の生物種の配列を含む格子点は，黒色で示した．主成分分析で求めた初期ベクトルを持つ格子点への，配列の帰属の結果を PCA として示した．格子点の色付けの方法は，学習後の SOM と同様に行っており，単一の生物種の配列のみが帰属した格子点の数は極端に少なく，殆どが黒色の領域である．（B）100 kb の 4 連塩基を SOM 解析した際の，連続塩基ごとの頻度分布の例．ゲノムサイズの大きな生物種の領域は，大文字のアルファベットで示した．連続塩基ごとの頻度分布図においては，使用頻度の高い順から赤・白・青と表示している．詳しくは，原著論文（Abe et al. (2003)）を参照されたい．**

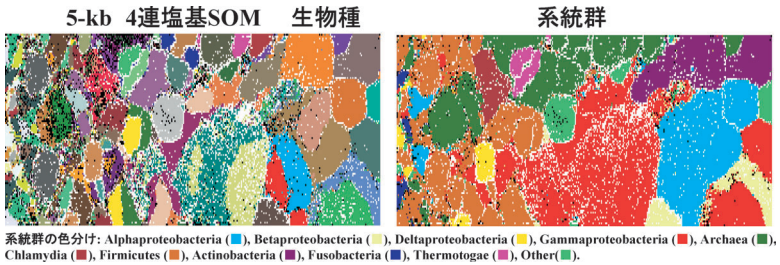

図 2. **配列既知なバクテリア 65 種の 5 kb 配列について 4 連塩基の出現頻度について SOM を行ったときの，バクテリア生物種ならびにその 11 の主要系統群への分布図．色付の格子点は 1 つの生物種（ないしは系統）だけからなるものであり，複数の生物種（ないしは系統）の配列を含む格子点は，黒色で示した．約 90%の 5 kb 配列が正しい系統群に分離しているので，生物種名が不明の混合ゲノム由来の断片配列をこの SOM へマップすることで，生物種を反映した分離が期待できる．存在する生物種の多様性やそれらの系統の推定が可能になる．**

種の細菌由来の配列の全体を 5 kb に断片化し，4 連塩基の出現頻度について SOM マップを作成した．難培養性の微生物を対象にした場合，未知の生物種である可能性が高く，どの生物系統に最も近いのかを推定することが重要になる．その目的で，上述の 65 種の細菌のゲノム配列について，主要な 11 の細菌の系統群への分類の様子も解析しているが，約 90%の配列が正しい系統を反映している．多種類のゲノム DNA が混在している環境試料由来の多数の断片配列を，図 2 の地図上へマップすることで，どの程度の種類でどの生物系統のゲノムが，どのような量比で混在していたのかを推定可能になる．将来的には，DNA データベースに収録された，生物種が判明している全てのゲノム配列で SOM を作成しておき，未知試料の混合ゲノムの配列をマップ可能にする予定である．

7. まとめ：統計数理研究者のゲノム配列解析への参画が重要と考えられる現状

配列の相同性検索やそれを基礎にした生物系統学を代表例として，統計数理学はゲノムの配列解析において重要な貢献をしてきた．本稿で紹介した SOM 解析も，統計数理学の新規な視点でのゲノム研究への貢献の例となる．初期ベクトルの設定に用いた主成分分析は，従来からゲノム配列の解析に用いられてきており，線形写像を基礎にしている．一方，SOM は非線形写像を基礎にしており，似かよった情報が二次元的に自己組織化する（おのずと組織化する）．高次元のデータをクラスタに分けて視覚化するための有効なツールであり，高次元空間で近接するデータを，高次元の情報を有効にいかしながら二次元的に近接させる．高いクラスタリング力を基礎に，生データ間での類似度からカテゴリーを同定し，個々のカテゴリーに反映される因子を捜すことも可能にしている．

筆者らが，ゲノム配列が解読された約 200 種の生物種の 5 連塩基頻度の SOM を作成する場合，国立遺伝学研究所のスーパーコンピュータの 32 CPU を用いた並列計算でも，一週間程度の長大な計算が必要になる．1,024 次元空間の数十万点の SOM であり，5 万個程度のニューロンを設定している．DNA データベースに収録された全てのゲノム配列を対象にした場合には，さらに大量なデータになる．また生物の機能に重要なシグナル配列を探索するには，6–10 連塩基へと解析を進める必要がある．6 連塩基（4,096 次元），7 連塩基（16,384 次元），8 連塩基（65,536 次元）へと連塩基の長さを単純に増やしていけば，高性能な計算機といえども実質的に計算が不可能になる．しかし，様々な統計数理的な手法を導入すれば，生命科学の研究者が遺伝シグナルとして興味を持っている，6–10 連塩基の解析も可能に思える．また，統計数理学にも新たな分野の展開が見られるかもしれない．

参 考 文 献

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2002). A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: Self-organizing map of oligonucleotide frequency, *Genome Informatics Series*, **13**, 12–20.

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2003). Informatics for unveiling hidden genome signatures, *Genome Research*, **13**, 693–702.

池村淑道 (2002). 配列既知の全ゲノムを対象にしたゲノム個性の解析，学術月報, **55**(12), 12–15.

Kanaya, S., Kudo, Y., Nakamura, Y. and Ikemura, T. (1996). Detection of genes in *Escherichia coli* sequences determined by genome projects and prediction of protein production levels, based on multivariate diversity in codon usage, *CABIOS*, **12**, 213–225.

Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. (1999). Studies of codon usage and tRNA genes

of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-speciffic diversity of codon usage based on multivariate analysis, *Gene*, **238**, 143–155.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H. and Ikemura, T. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): Characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome, *Gene*, **276**, 89–99.

金谷重彦, 阿部貴志, 木ノ内誠, 池村淑道 (2003). **ゲノム配列に潜む特徴的なサイン**(Genome Signature)**から見た生物多様性,『ゲノムからみた生物の多様性と進化』,** 58–64, **シュプリンガー・フェアラーク東京, 東京**.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, **43**, 59–69.

Kohonen, T. (1990). The self-organizing map, *Proceedings of the IEEE*, **78**, 1464–1480.

コホネン T. (1993). **『自己組織化と連想記憶』(中谷和夫 監訳), シュプリンガー・フェアラーク東京, 東京**.

Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. (1996). Engineering applications of the self-organizing map, *Proceedings of the IEEE*, **84**, 1358–1384.