統計数理(2008)
第 56 巻 第 1 号 67–79

特集「分子進化と統計科学」
[原著論文]

トーラス型学習領域を持つ自己組織化マップ法を用いたゲノム配列の系統関係

洞田 慎一[1] ・池村 淑道[2] ・湯川 哲之[3]

(受付 2007 年 7 月 13 日；改訂 2007 年 9 月 13 日)

要　　旨

ニューラルネットワーク・アルゴリズムを応用した自己組織化マップ法(SOM)によるクラスタ分類をゲノム配列に応用し，クラスタ間の相対関係からゲノム間の系統関係を議論する．SOM をゲノム配列解析に適用すると，断片配列の連文字頻度分布のクラスタ分類から，大量の入力データに対しても効率よく系統群の分類が行えることが知られている．本研究では，分類されたクラスタの相対関係から入力したゲノム配列の間に存在する系統関係を考察する目的のために，クラスタ間の相対関係を評価可能なトーラス型 SOM を用いた解析方法を紹介する．これにより，クラスタ間の相対関係が明確化し，近接するクラスタの距離解析からゲノム配列の系統解析が可能となる．実際にインフルエンザウィルスに対する推定を SOM 解析により評価した．

キーワード：SOM，自己組織化マップ，バイオインフォマティクス，ゲノム配列，連文字頻度解析．

1. はじめに

SOM は，ニューラルネットワークを応用し，ニューロンの学習機構を用いた入力データの特徴抽出を効率的に行うデータマイニング手法として，コホネンにより提案されている(Kohonen, 1982, 1984, 1990, 1995)．この方法は，入力データと同じ自由度を有するニューロンに入力データの傾向を学習させ，自己組織化マップと呼ばれるデータの特徴を反映したクラスタ分類により，特徴をより明確化する手法である．学習及び解析の手法はまったく分類の施されていない未知のデータを扱うことが可能で，データの量や構造といった性質に依存しないことから，様々な方面に応用が可能である．

ゲノム配列の解析に応用した事例として，データの入力順序に依存しない一括学習を行う，教師なしバッチラーニング SOM (BL-SOM)法を利用した阿部らの一連の解析(Abe et al., 2003, 2005; 阿部 他, 2004; Kanaya et al., 2001)が知られている．彼らは，ゲノム配列の特徴抽出を行うため，断片配列のオリゴヌクレオチド連文字頻度分布を入力データとして，断片配列内で特徴的な塩基文字列の連文字頻度情報を抽出している．つまり，入力した生物種を特徴づける単語の使用頻度による個性をクラスタとして発見することにより，ゲノム配列に含まれる生物種

[1] 総合研究大学院大学 葉山情報ネットワークセンター：〒 249–0193 神奈川県三浦郡葉山町湘南国際村
[2] 長浜バイオ大学 バイオサイエンス学部：〒 526–0829 滋賀県長浜市田村町 1266 番地
[3] 総合研究大学院大学 葉山高等研究センター：〒 249–0193 神奈川県三浦郡葉山町湘南国際村

固有の情報を抽出可能であることを示した.

ただし,どのような情報が特徴抽出に重要な役割を果したのか,あるいは生物種を分離した原因を特定するという問題は,クラスタ分類後の解析を必要とする.そこで,本研究では,SOM に現れる生物種クラスタについて,その系統関係がどのように反映されるかを,各クラスタの相対関係や構造に注目して解析を行った.特に,クラスタの近接関係や大きさといった相対的な情報は,入力したデータの相対的な関係を反映していると期待できる.例えば,各クラスタの位置はそれぞれの類似度に関係し,隣接するクラスタは遠く離れたクラスタよりも近い性質を持っていると期待できる.さらに,クラスタの内部に見られる細かなサブクラスタ構造は,データがさらに細かな構成要素から成り立っていることを意味する.つまり,各クラスタの相対関係や構造を明確にできれば,SOM を用いたゲノム配列解析において,生物種を分類するのと同時に,系統関係や,さらなる内部構造を発見することが可能で,昨今の大量情報による広範囲のゲノムデータの整理を行うツールとして,SOM が重要な役割を担うと期待できる.

以上の目的をふまえて,SOM に現れるクラスタの相対関係に注目して,ゲノム配列解析を行う.クラスタ同士の相対関係を議論するために,各クラスタが等方的な関係を保つよう,従来の長方形型のニューロン結合規則ではなく,四方の辺をそれぞれ周期的境界条件で接続したトーラス型のニューロン配置を用いて解析を行った(Horata et al., 2005a, 2005b).

前述したように,SOM はニューラルネットワークを応用したデータマインニング手法であり,入力データ,ニューロン,出力データの 3 つの層から構成される.クラスタ分類の要となる入力データの傾向を学習するニューロン層では,ニューロンがそれぞれの近傍の情報を元に学習を行うため,格子グリッド状にニューロンが配置される.各ニューロンの結合規則は,学習の結果である出力層におけるクラスタ構造に影響することがある.例えば,四方に境界を持つ長方形型の格子グリッド状に配置する場合,境界が特別な役割を果たしてしまい,クラスタの相対関係が不明瞭になる恐れがある.一方,境界ニューロンを持たないトーラス型のニューロン結合規則では,ニューロンへの学習効果が各方向から均等に行われるため,各クラスタの相対的な情報を保持するのに有利な方法であると考えた.

出力層における各クラスタの相対関係を解析するために,各クラスタを構成するニューロンから代表値を作成し,クラスタ間の相対的なユークリッド距離を測定した.さらに,この幾何情報をゲノム間距離と考えて系統樹作成を試みた.系統樹作成は,種の相対的相違を比較検討する上で有効であり,客観的に作成するにはいくつかの方法(近接接合法(Saitou and Nei, 1987),最尤法(Felnsenstein, 1981; Hasegawa and Yano, 1984)が知られているが,ここでは SOM に現れるクラスタの情報が,生物間の相対的な情報を保持しているかどうかに注目して解析を行った.つまり,ゲノム配列の違いを SOM から得られる数値データに置き換え,得られたクラスタの相対関係から,その系統関係を比較しようとするものである.

本論文では,SOM を周期的境界条件を持つトーラス面に写像した場合のゲノム配列解析方法を説明し,実際にこの手法をインフルエンザウィルス解析に応用した事例を示す.

2. トーラス型 SOM のゲノム配列解析への応用

SOM は,入力データを配置する入力層と,データの学習を行うニューロン層,その結果をクラスタ分類として可視化する出力層から構成される(図 1).特に,ゲノム配列の解析に有効と考えられている,教師なしバッチラーニング SOM (BL-SOM)法を利用したアルゴリズムの手順(Kanaya et al., 2001)は,次のようにまとめることができる.

(1) 解析したい M 個の入力データの集合 $\{\vec{x}_m, m=1,\ldots,M\}$ を考え,それぞれを入力層に割り当てる.教師なしバッチ型学習(BL-SOM)法を行う場合,入力層は任意次元のベクトル $\vec{x}_m$

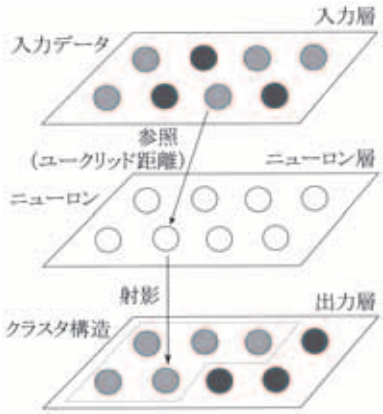

図 1.　SOM アルゴリズムにおける入力層とニューロン層の関係の概略図．入力層では入力データが配置され，それぞれの入力データ $\vec{x}$ を丸印 (●) で示す．色の違いはそれぞれの値の違いを表現した．ニューロン層では，各ニューロン (丸印 (○) で示す) が配置され，入力層の各データとユークリッド距離を比較し，最も距離の短いものが参照され，出力層に射影される．そのため出力層では，ニューロンに応じて，入力データのそれぞれの相違が並べ直され，クラスタ構造を見いだすことができる．

から構成される $M \times \dim(\vec{x})$ 次元行列である．

(2)　ニューロン層として $I \times J$ 個のニューロンを格子グリッド $\{(i,j) = (1,1) \cdots (I,J)\}$ に配置する．各ニューロンに入力データ $\vec{x}_m$ に対するリファレンスとしての重みベクトル $\vec{w}_{i,j}$ を付与する．

(3)　入力データ $\vec{x}_m$ と重みベクトル $\vec{w}_{i,j}$ のユークリッド距離を比較し，最も近い入力データをニューロンと関係づける．

(4)　重みベクトル自身は学習により変化し，入力データとの関係付けが収束した状態として，出力層を作成する．

実際の各ニューロンにおける学習は，近傍領域にあるニューロンに関係づけられた入力層のデータを用い，重みベクトルを式(2.1)にて逐次更新することで行われる．

$$\vec{w}_{i,j}^{next} = \vec{w}_{i,j}^{prev} - \alpha \frac{\sum_{i',j' \in S(\beta)} \vec{x}_{i',j'}^{m}}{\sum_{i',j' \in S(\beta)} 1} \tag{2.1}$$

ここで，α は，学習の程度をコントロールする学習パラメータであり，β は，学習で参照するニューロンの近傍領域 $S(\beta)$ を定義する領域パラメータである．近傍領域 $S(\beta)$ は，具体的には，$i-\beta \leq i' \leq i+\beta, j-\beta \leq j' \leq j+\beta$ の条件を満たす格子グリッド点 (i',j') である．パラメータ α, β を，学習の収束のために減少させながら学習を行わせる．学習を繰り返すことで，各ニューロンの近傍には類似した重みベクトルを持ったニューロンが集中し，出力層には入力データの特徴を反映したクラスタ構造が現れる．クラスタの判別には，各ニューロンの重みベクトル間の距離を比較した距離行列としての U 行列を用いることも可能である (Ultsch, 2005)．学習の過程を段階的に追っていくと，性質が明瞭に異なるクラスタは，重みベクトルが入力データの傾向を学習するにつれて，出力層での相対的な位置が離れていくような効果 (Flanagan, 2000) を確認することができる．例えば，あるパラメータで記述される関数式から入力データを作成すると，クラスタ構造はパラメータによって連続し，パラメータの値によって性質が大きく異なるクラスタ間の相対位置は大きく離れることが確認できる．演繹すれば，クラスタ構造は入

力データ自体の特徴だけではなく，それぞれの特徴の相対関係や内部構造といった情報も反映していると考えられる．

ただし，クラスタ構造の相対関係を解析する為には，学習の各段階でのニューロン近傍領域の取り方に注意が必要である．なぜならば，境界のある長方形型格子グリッドでは，式(2.1)におけるニューロンの近傍領域 $S(\beta)$ は，境界を越えて広がることはないため，ニューロンの学習機会が境界近傍と，中心近傍では均一ではない．格子グリッド点 $i \leq \beta, i \geq I-\beta$ あるいは，$j \leq \beta, j \geq J-\beta$ の領域では，参照できない格子グリッド点が存在する．この学習近傍の差は，クラスタの相対的な情報に影響する可能性がある．例えば，性質の異なったクラスタは境界方向へ弾き出されるが，境界を越えて押し出されることがないため，格子グリッドの境界に張り付いて安定してしまう状況が確認できる．

ゲノム解析においては，生物種間の系統関係などの相対的な情報はクラスタ同士の相対関係に影響すると期待できることから，クラスタ間の相対関係を正しく評価する事は重要な問題である．そこで，ニューロン層の格子グリッド境界を解消するように，長方形の向かい合う辺を周期的境界条件にて接続したトーラス型格子グリッドの結合規則を作ることが適当と考えた．この場合，学習の各段階での近傍領域は，境界を挟んで周期的に隣接するニューロンを選択できるようになるため，ニューロンの学習機会の不均一性は解消し，クラスタの相対関係に境界上ニューロンが特別な影響を与えないと考えられる．トーラス上に配置したニューロンの重みベクトルは，次の周期的境界条件を満たす．

$$\vec{w}_{-i,-j} \cong \vec{w}_{I-i,J-j} \tag{2.2}$$

実際には，SOM アルゴリズム上ではニューロン層の結合方法は自由に選択できるため，他に球体といった格子グリッド空間のトポロジーも提案されている(Nishio et al., 2004)．

このようなニューロン格子グリッドの取り方の他，入力データをどのように準備するかという問題も重要である．すでに阿部ら(Abe et al., 2003; 阿部 他, 2004)により，ゲノム配列から入力データを作成する手法が提案されている．この手法は，ゲノム配列を長大な 1 次元文字配列と考えて，あるウィンドウ単位毎に断片配列を切り出し，その断片配列における部分連文字列の出現頻度を入力データとするものである．オリゴヌクレオチドの文字数を W とすると，4 つの塩基文字の組合せで発生する組合せの数から，入力ベクトル $\vec{x}$ は，$\dim(\vec{x}) = 4^W$ 次元のベクトルとして扱われる．例えば，“ATGGATAGCGTA” という断片配列を 2 連塩基にて数えた場合の連文字頻度による入力ベクトル $\vec{x}$ は，次のように与えられる．

$$\begin{cases} x(AA)=0, x(AT)=2, x(AG)=1, x(AC)=0, \\ x(TA)=2, x(TT)=0, x(TG)=1, x(TC)=0, \\ x(GA)=1, x(GT)=1, x(GG)=1, x(GC)=1, \\ x(CA)=0, x(CT)=0, x(CG)=1, x(CC)=0 \end{cases} \tag{2.3}$$

このような断片配列のオリゴヌクレオチド連文字頻度分布からの解析は，異なった生物種に由来する断片配列を比較するのにアライメントなどの操作を必要とせず，培養が困難な新規性の高い微生物種や，複数の生物種が混在する試料に由来した断片配列の分類にも有効な手法である．本研究での解析も同じ手法を用い，オリゴヌクレオチド連文字頻度分布ベクトルを対象とする生物種のゲノム配列のそれぞれについて計算し，入力層となる数値行列を作成する．

次に，学習の初期条件にあたる，各ニューロンの重みベクトルの初期値を与える必要がある．境界のある長方形型格子グリッド平面の場合について，金谷ら(Kanaya et al., 2001)により，主成分分析を基本とした手法が提案されている．この手法をトーラス型格子グリッドの境界条件を満たすように拡張する．トーラスとは，長方形の向かい合う辺を周期的境界条件にて接続し

た図形であり，中央に穴の開いたドーナツ型の立体の表面として描かれる．これを平面に展開すると，2 本の直行する軸を持った境界のない 2 次元平面と考えることができる．展開した平面グリッドを再びドーナツ型の表面に貼り付けると，各格子グリッド点に割り当てる初期ベクトルは，ある任意の 1 点に割り当てた初期ベクトルを各方向へ回転させることで算出できることがわかる．格子グリッドにおける直交する 2 軸は，ドーナツ型の立体表面を構成する 2 つの円(α-cycle，β-cycle)にそれぞれに対応している．それぞれの直交する軸を，境界のある場合と同じように第 1 主成分ベクトルと第 2 主成分ベクトルに対応させようとすれば，一つの方法として，第 1 主成分ベクトルと第 2 主成分ベクトルを互いの周りに回転させ，周期的な 2 軸を便宜的に作成することで周期的境界条件の解決が可能である．そこで，トーラス型格子グリッドの任意の一点 $(i,j)=(1,1)$ に，第 1 主成分ベクトル $\vec{p}^{\,1\mathrm{st}}$ と第 2 主成分ベクトル $\vec{p}^{\,2\mathrm{nd}}$ を割り当て，2 軸(α-cycle，β-cycle)に関連させた基底ベクトル $(\vec{c}^{\,\alpha-\mathrm{cycle}}(1,1),\vec{c}^{\,\beta-\mathrm{cycle}}(1,1))$ と考える．

$$\begin{cases} \vec{c}^{\,\alpha-\mathrm{cycle}}(1,1)=\vec{p}^{\,1\mathrm{st}}, \\ \vec{c}^{\,\beta-\mathrm{cycle}}(1,1)=\vec{p}^{\,2\mathrm{nd}} \end{cases} \tag{2.4}$$

隣接する格子グリッド点での基底ベクトルは，互いのベクトル周りに回転させることでトーラス格子グリッドが再現できる．具体的には，出発点となる格子グリッド点 (i,j) にて基底ベクトル $\{\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j),\vec{c}^{\,\beta-\mathrm{cycle}}(i,j)\}$ について，互いに直交する法線ベクトル $\vec{n}_{i,j}$ をベクトル外積演算にて求め，基底ベクトルを法線ベクトルを用いて互いのベクトル周りに回転させる．

$$\begin{cases} \vec{n}_{i,j}=\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j)\times\vec{c}^{\,\beta-\mathrm{cycle}}(i,j), \\ \vec{c}^{\,\alpha-\mathrm{cycle}}(i+1,j)=\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j)+\vec{n}_{i,j}, \\ \vec{c}^{\,\beta-\mathrm{cycle}}(i+1,j)=\vec{c}^{\,\beta-\mathrm{cycle}}(i,j), \\ \vec{c}^{\,\alpha-\mathrm{cycle}}(i,j+1)=\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j), \\ \vec{c}^{\,\beta-\mathrm{cycle}}(i,j+1)=\vec{c}^{\,\beta-\mathrm{cycle}}(i,j)+\vec{n}_{i,j} \end{cases} \tag{2.5}$$

これを各格子グリッド点にて繰り返し実行することで，トーラスを直交グリッドに展開した格子の各頂点上の基底ベクトルの集合 $\{\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j),\vec{c}^{\,\beta-\mathrm{cycle}}(i,j)\}$ を定義できる．

トーラスの格子グリッドの大きさ (I,J) は，第 1 主成分，第 2 主成分それぞれの方向についての分散量の大きさに比例するように選ぶ．具体的には，基底ベクトル $\{\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j),\vec{c}^{\,\beta-\mathrm{cycle}}(i,j)\}$ を用いて，初期重みベクトルを次のように設定した．

$$\vec{w}(i,j)=\vec{x}+5\sigma^{1\mathrm{st}}\vec{c}^{\,\alpha-\mathrm{cycle}}(i,j)+5\sigma^{2\mathrm{nd}}\vec{c}^{\,\beta-\mathrm{cycle}}(i,j) \tag{2.6}$$

$\vec{x}$ は入力データの平均値であり，$\sigma^{1\mathrm{st}},\sigma^{2\mathrm{nd}}$ は，主成分第 1 軸と第 2 軸に対する分散である．式(2.6)の重みベクトルは式(2.2)の周期的境界条件を満たす．ただし，この手法では初期重みベクトルが周期的境界条件を満たすものの，初期状態において入力データと関連づけられるニューロンはトーラス面の一部にトラップされている可能性がある．そこで乱数で初期重みベクトルを作成する場合と比較したが，どちらもクラスタ分離が可能で，各クラスタ間の相対関係を解析しても結果は変わらなかった．ここで提案した式(2.6)の方法を用いた際に，入力データの初期マッピングには偏りがあったとしても，それは学習により解決が可能である．出発となる初期条件を固定する場合，常に計算で現れるクラスタ位置の絶対的位置を固定できる．

ゲノム配列から作成した入力データを元に平面の場合と同様に学習を行う．各ニューロンには式(2.1)にある学習を試行させる．パラメータ α,β は，学習の試行回数 t の関数として次のように変化させた．

$$\alpha=\max(0.01,\alpha_0(1-t/T)),\beta=\max(1,\beta_0-t) \tag{2.7}$$

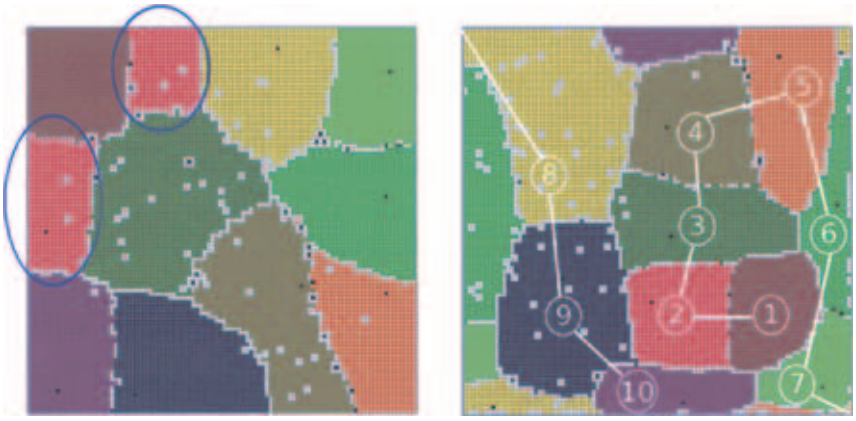

図 2. 世代毎に塗り分けたランダム配列マップ結果(左:プレーン，右：トーラス)．色の違いはそれぞれの世代を示す(1 世代から 10 世代)．世代とクラスタの対応は，右図の図中の数字が各世代の対応を示している．右図の図中の数字とそれを接続する線は，それぞれのクラスタの重心距離の最も短いものを選択し，結合したもので，各世代の接続関係を示す．SOM 数値解析は，初期パラメータを $T=400, \alpha_0=0.95, \beta_0=I/2$ と設定して，3 塩基連文字頻度分布を入力データとして解析を行った．

T は総学習回数であり，α_0, β_0 は適当な初期値である．本論文では，これらの初期パラメータを $T=400, \alpha_0=0.95, \beta_0=I/2$ に設定して，3 塩基連文字頻度分布を入力データとして解析を行った．塩基連文字列の長さは，データの解像度や情報量に関係すると考えられるが，基本的には，いずれの連文字頻度においてもクラスタを見出すことは可能で，データの解像度と計算時間を照らし合わせて，これらのパラメータを選択した．

3. SOM における生物種系統関係

上述した方法，つまりニューロン層をトーラス型格子グリッドに配置することがクラスタ間の相対関係を調べるのに適するかどうかについて，人為的に作成したランダムな配列を作成し，学習の様子とクラスタ間の相対関係を調べた．特に，比較的類似したゲノム配列間の系統関係を再現できるかどうかを見るため，作成するサンプルに類似性を持たせた．基本となる入力配列から各塩基文字に対してランダムに塩基置換を起こさせて，10 世代の 100-kb サンプル配列を作成した．最初の世代を作成するための母配列は，モノヌクレオチド出現頻度 $\{a_i, i=1,2,3,4\}$ を初期値として与え，その構成比率に合うように人為的なゲノム配列を乱数を用いて作成する．その配列を中央値として，それぞれ塩基置換の確率を次のように設定した．

$$P=\frac{1}{2\pi}e^{-\sum_i^4 \frac{(x_i-a_i)^2}{v_i}} \tag{3.1}$$

分散量 $\{v_i, i=1,2,3,4\}$ は塩基置換により集めるサンプルのばらつきを調整するパラメータとして適当な値を入力する．塩基置換を 1000 回繰り返し起こさせ，塩基文字のばらつきが正規分布するように，サンプル配列を作成する．同時に，クラスタ間の相関関係を調べるため，無作為に選んだ一本を次の世代の母配列として逐次繰り返しサンプルを作成する．図 2 は，作成した 10 世代のランダム配列サンプルについて，SOM 解析を行った結果である．

図 2 の左図は，境界のある長方形型格子グリッドを用いた場合の解析結果であり，右は境界がないトーラス型格子グリッドを用いた場合の解析結果である．いずれのマップについても，

世代に対応したクラスタ構造を見いだすことができる．ただし，長方形型格子グリッドの場合，サンプルの性質によって，ある世代のクラスタが二つに分離する状況が見られるが(図 2 左，丸く囲んだクラスタ)，トーラス型格子グリッドの場合にはそのようなクラスタの分離は起きておらず，クラスタ間の関係を正しく評価できる可能性があることを数値的に示している．各ニューロンの学習差が生じないことからも，クラスタ間の相対的な情報の比較を行う場合にはトーラス型格子グリッドのように，境界を持たない配位を選択するのが適当と考えられる．

クラスタ間の相対関係を調べるために，クラスタ間の重心距離に注目した．クラスタを構成するニューロンには学習後の重みベクトル $\vec{w}_{ij}$ が情報として残されているため，これを用いてクラスタの重心位置を計算することができる．

$$\vec{w} = \frac{\sum_{i,j\in\mathrm{cluster}} \vec{w}_{i,j}}{\sum_{i,j\in\mathrm{cluster}} 1} \tag{3.2}$$

つまり，最も重心位置に近い距離にあるニューロンをクラスタの重心ニューロンとした．クラスタを構成するニューロンについて，どのニューロンからも相対距離が平均的に近いニューロンを抽出し，クラスタの重心位置とするものである．重心ニューロンは，クラスタの特徴を平均的に反映しているであろうと推測できる．人為的に作成したランダム配列の結果について，重心ニューロンをそれぞれの世代のクラスタについて計算すると，その世代の母インプット配列である前世代のサンプルの一つと一致する．サンプルの構成方法から考えると，この結果は妥当である．従って，SOM に見いだされるクラスタは，重心位置にその性質の代表値をもち，重心ニューロンの周囲に類似したニューロンが集合して構成されていると考えることができる．各クラスタ間の重心を計算し，それぞれの重みベクトルの距離を計算することで，各クラスタ間の重心距離行列を作成できる．ランダム配列の場合，連続する世代のクラスタ間の距離が最も近くなり，距離の近いクラスタは，それぞれ近接する．図 2 右の図中に数字と線として示したように，重みベクトルから算出した距離に基づいて，クラスタ間の系統関係を推定することができる．つまり，それぞれのクラスタに隣接するクラスタ間の距離行列を作成し，これとクラスタの近接情報から親子関係を見出し，系統樹としてまとめることができる．

重心ニューロンの導入によって，クラスタ間の相対関係を把握できる可能性が数値的に示されたことから，これを実際の生物種のゲノム配列解析に適用した．ここでは，実際の生物種として，インフルエンザウィルスの SOM 解析を行った．このサンプルのゲノム配列データは，サンプリングした時期や亜種，地域毎の付加情報を含め，比較的広範囲かつ正確に集められており，それぞれの系統関係を掴み易い為，SOM を用いて多様性や系統関係を調査するに最も適した対象といえる．2004 年までに登録されているすべてのゲノム配列，11585 サンプルについて SOM 解析を行った(図 3)．

インフルエンザウィルスは，A 型，B 型，C 型の 3 種類が存在し，RNA ウィルスとして，8 本の RNA ゲノム分節を持つ(ただし C 型は 7 本の分節をもつ)．また，ウィルスの表面には，スパイク状の糖タンパクが突出し，A 型ウィルスはスパイクタンパクの型の違いにより H1 型から H15 型，N1 型から N9 型のサブタイプに区分される．3 種類存在するインフルエンザ型に応じてクラスタの塗り分けや，ゲノム分節，H1 型から H15 型のサブタイプについて塗りわけを行うと，それらは SOM 解析により，はっきりと見分けることができる．つまり SOM に現れるクラスタには，その内部に複雑なサブクラスタを含み，サブクラスタもまた生物学的な情報を含んでいることを意味している．例えば，A 型ウィルスクラスタの内部には，それぞれのゲノム分節毎のクラスタが存在し，さらにタンパク質の違いによりサブタイプのクラスタを有する(図 4)．ただし，サブクラスタの包含関係はそれぞれの構成に依存するため複雑である．

ウィルスが収集された地域による分類を調べるために，H5 型トリインフルエンザウィルス

図 3. インフルエンザウィルス (11585 サンプル) のゲノム配列について自己組織化マップした結果．B 型，C 型については，図中では (B，C) とまとめ，A 型については，ゲノム分節 (PB1，PB2，PA，HA，NA，NP，M，NS) による塗り分けを行った．図中の文字の位置にそれぞれのクラスタが存在する．SOM 数値解析は，初期パラメータを $T = 400, \alpha_0 = 0.95, \beta_0 = I/2$ と設定して，3 塩基連文字頻度分布を入力データとして解析を行った．

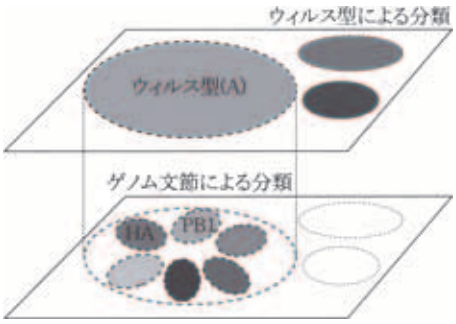

図 4. サブクラスタの構造．図 5 に示した SOM の結果は，A，B，C 型のそれぞれに分類でき，さらに A 型の例で示したように，ゲノム分節のようにさらに細かなクラスタ分類が可能であることを示している．

について，HA ゲノム分節での分類状況を国別に塗り分けた結果を図 5 に示した．この結果から，ウィルスが移動した場合に，そのゲノム情報を少しずつ変化させていることが予測される．トリインフルエンザは，主に渡り鳥が伝搬し，生き残ったウィルスが地域を越えて伝わっていくことから，その効果が地域間での変化を生んでいると期待でき，その影響を SOM においても見出すことができたと考えられる．また，トリやウマなどに感染するウィルスのタイプは国別で高い相関が現れるが，ヒトに感染する種類の場合，国別では明確にクラスタ分離するのは困難である．

さらに，経年変化を調べるため，H3N2 型ウィルスの HA ゲノム分節について，1983 年から 2001 年にかけての変化を経年別にクラスタの塗り分けを行った．図 6 は，サンプル年数を 1 年

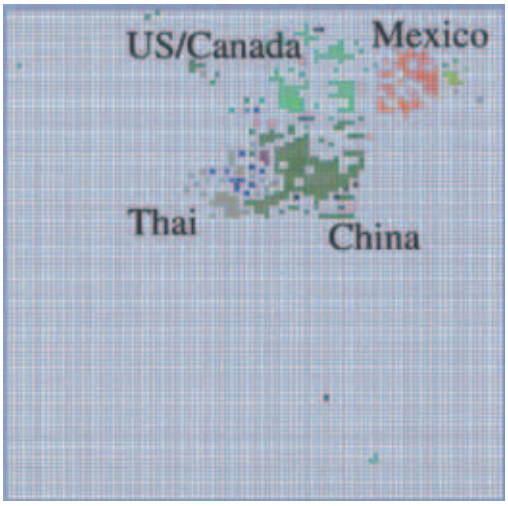

図 5.　図 3 に示した 11585 サンプルでの SOM 数値計算結果のうち，特に H5 トリインフルエンザウィルスに注目し，サンプル国別にクラスタを塗り分けた結果．便宜的にメキシコ (橙)，アメリカ，カナダ (黄緑)，中国 (緑)，タイ (灰) とした．

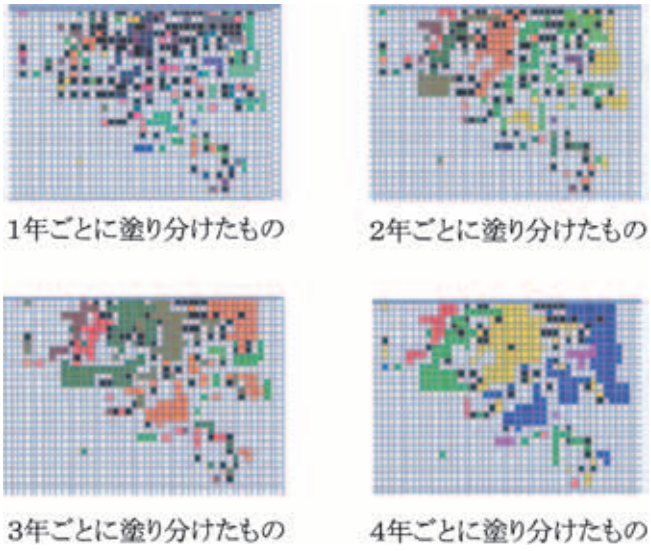

図 6.　図 3 に示した 11585 サンプルでの SOM 数値計算結果のうち，H3 型 HA 鎖に注目して，それぞれのサンプリング年代による分類 (1983–2001) を行ったもの．色の違いが年代の相違を表す．例えば，1 年毎に塗り分けた場合，それぞれの年に採取されたそれぞれのサンプルを示し，2 年毎では，サンプル採取の 2 年間の間に採取されたグループとしたもの．

毎から 4 年毎にクラスタを塗り分けたものである．定量的な比較は 1 年毎ではやや困難ではあるが，おおよそ 2 年毎から 3 年ごとに塗り分けた場合からクラスタの構造が明確になる．これは，インフルエンザウィルスのゲノムの明瞭な塩基置換がおおよそ 2 年から 3 年の間で起き

表 1. H3 型 HA 鎖クラスタ重心ゲノム配列サンプル.

年	ACCESSION 番号	Definition
1983	AF008906	Tonga/23/85
1989	L20114	Singapore/12/89
1989	L20115	Singapore/12/89
1992	Z46408	HongKong/2/94
1995	AF008725	Nanchang/933/95
1998	ZF534036	Mendoza/133/99
1998	AF534032	BuenosAires/M6/99
2001	AY138518	Ningbo/17/2002

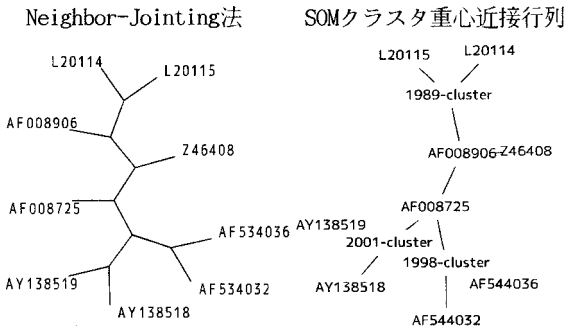

図 7. H3 型 HA 鎖における無根系統樹の推定結果. 図 6 で得られた 2 年毎に塗り分けたクラスタの重心から推定した.

ていることを示唆している. 特に, 約 2000 塩基対からなる HA ゲノム分節について, その非同義置換率は 1.41×10^{-3} 変異/塩基/年であること(吉倉, 2001)から, この数値計算からの示唆は妥当なものであると考えられる. また, このような経年変化がみられる状況は H5, H9 型といったサブタイプでも見られ, この場合, 上述したように地域別の変化とも対応している. さらに, ランダム配列の場合と同様にクラスタ重心間の距離行列と, クラスタの近接関係からそれぞれのクラスタの系統関係を調べた. 特に, 時系列変化に注目するため, H3N2 型ウィルスの HA ゲノム分節について, 1993 年から 2001 年にかけての系統関係に注目した. 上述したように, 2,3 年毎にクラスタを追いかければ, 経年変化を識別できることから, クラスタを 2 年毎に塗り分けした場合について, それぞれの重心位置を求めた. 重心位置に関連するサンプル配列の具体的な情報は表 1 に示す. 一つの重心ニューロンに複数本の入力サンプルが割り当てられている場合は, それらを列記した. それぞれのサンプルのゲノム配列をアライメントし, 近隣接合法により系統樹を推定したものと, クラスタ重心ニューロンの重みベクトル距離行列から推定した系統関係を図 7 に示す. SOM 解析では各クラスタの重心ニューロンについての距離行列を求め, 近接するクラスタのうち距離の短いものから結合していく. サンプルにも依存はするが, 近接するクラスタのうち距離が最も近いものは, 他のクラスタとの距離と比較して極端に異なるため, 区別することが可能である. 一方, 接続が分岐する状況では, それぞれの大きさがほとんど区別できない. 実際の手順としては, それぞれの重心点において距離を比較し, 2 値的な閾値を算出して判別し, それを基に, 距離の短いものから順に結合していく. また, インフルエンザウィルスではサンプル年代が特定でき, 接続の相互関係をある程度見分け

ることは可能であることから，手順をルール化する際のチェックとなる．各枝の接合を比較すると，系統関係の概略は両者で一致しており，SOM を用いても近隣接合法と同様に系統関係を見出すことが可能であるように見える．これは，連文字頻度情報に基づいた特徴を SOM で抽出した結果，学習後の各ニューロンの重みベクトルは分類された入力ベクトルに対して，ベクトルの平均値を表しており，重心ベクトルはそのクラスタを構成する全入力ベクトルの平均ベクトルに近いと考えられ，分類されたクラスタの連文字頻度情報の特徴を平均的に反映しているためと考えられる．より詳細に関係を調べるためには各枝の変異距離を比較する必要があるが，SOM を用いて入力した生物種のそれぞれの相対関係を議論することができることを示している．

4. 考察

トーラス型 SOM 解析を行うことで，出力層に出現するクラスタは，それぞれ入力データの相対関係が保持された状態でクラスタが形成されていることが示された．また，学習の初期条件として必要となる各ニューロンの重みベクトルの初期値を与える一手法を提案したが，乱数で与えた場合と比較しても明瞭な分類ができる．初期値を固定した場合，出力層におけるクラスタの出現位置が固定され，解析が容易である．

ゲノム解析において，SOM を用いることでオリゴヌクレオチド連文字頻度情報から生物種への分類を行うことができるが，クラスタ分類が可能であった理由や，さらに生物学的情報を取り出すためには，クラスタの相対関係や構造に対する数値解析が欠かせないものとなる．本論文では，クラスタの特徴を重心ニューロンに代表させ，それぞれのニューロン間の重みベクトルの相対的な距離や近接関係を調べることで，クラスタの系統的な関係性を探ることが可能であることを示した．入力データとしてゲノム配列におけるオリゴヌクレオチド連文字頻度分布を用い SOM で特徴抽出した結果，学習後の各ニューロンの重みベクトルは分類された連文字頻度分布に対して平均値に近い値が選ばれることから，重心ベクトルはそのクラスタを構成する連続頻度分布の平均的な特徴を抽出していると考えられる．このため，クラスタ間の相違は，入力したゲノム配列の塩基置換の程度に関係し，クラスタを代表する重心間の相対的な距離関係は，ゲノム配列の置換回数の違いと類似した尺度を有していると期待できる．そのため，クラスタ間の相対距離や近接関係を議論するには，境界を取り除いてトーラスのように配置し，クラスタ間の相対関係を議論しやすくしておくことが必要であると考えられる．

どのクラスタが近接するかは，それぞれの相対的な情報により決まることから，おおよその系統情報が SOM により割り出せることを意味している．ただし，SOM により算出される系統関係は入力したオリゴヌクレオチド連文字頻度情報の相対的な関係のみであり，この方法は，SOM による分類を行った際に，さらに情報を捉むための一手法と考えるべきである．より正しい系統関係を導き出すには，リボソーム RNA の遺伝子のように，オルソロガスな配列が広範囲の生物種で知られている場合には，最尤法などで推定された祖先配列の情報を用いるなど，間接的なデータ取り扱いの手法が考えられる．最尤法などで推定される祖先配列からの連文字頻度情報を SOM 解析に加えることで，進化過程がより明確に推定可能になるであろう．

SOM では，ゲノム配列全体の情報から比較を行えるため，特定の遺伝子などに注目する必要性がなく，対象が定まっていない，あるいは未分類の大量な生物種の相対関係を探る方法論としては有効であると考えられる．SOM による系統関係抽出をより確かなものにするためには，今後，アルゴリズムのさらなる改良と，既存方法との比較を注意深く検討しなければならない．

参 考 文 献

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.

阿部貴志，金谷重彦，木ノ内誠，池村淑道(2004)．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的な統計的数理的手法，統計数理，**52**(1)，207–215.

Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M., Matsuura, Y., Tokutaka, H. and Ikemura, T. (2005). A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide range of eukaryotic genomes, *Proceedings of WSOM 2005*, 187–194.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences ― A maximum likelihood approach ―, *Journal of Molecular Evolution*, **17**, 368–376.

Flanagan, J.A. (2000). Neuron weight dynamics in the SOM and self-organized criticality, *IJCNN2000*, Volume 5, 39–44.

Hasegawa, M. and Yano, T. (1984). Maximum likelihood method of phylogenetic inference from DNA sequence data, *Bulletin of the Biometric Society of Japan*, **5**, 1–7.

Horata, S., Ikemura, T. and Yukawa, T. (2005a). Torus Self-Organizing Map for genome informatics, *Proceedings of WSOM 2005*, 235–242.

Horata, S., Ikemura, T. and Yukawa, T. (2005b). Torus Self-Organizing Map for genome informatics, *IPSJ Symposium Series Volume 2005*, No. 11, 71–78.

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.

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

Kohonen, T. (1984). Self-organization and assoxiative memory, *Springer Series in Information Sciences*, **8**, Springer-Verlag, New York.

Kohonen, T. (1988). An introduction to neural computing, *Neural Networks*, **1**, 3–16.

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

Kohonen, T. (1995). *Self-Organizing Maps*, Springer Series in Information Science, **30**, Springer-Verlag, New York.

Nishio, H., Wada, K., Wada, Y., Amin, M. and Kanaya, S. (2004). Suitability of spherical SOM for gene expression analysis, *Proceedings of RECOMB2004*, 79–80.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, **4**, 406–425.

Ultsch, A. (2005). U*F clustering: A new performant "cluster-mining" method based on segmentation of Self-Organizing Maps, *Proceedings of WSOM 2005*, 25–32.

吉倉 廣 監修(2001)．『ワンポイントウィルス学 ― Essentials of Medical Virology ―』(豊田哲也 編集)，南山堂，東京.

Proceedings of the Institute of Statistical Mathematics Vol. 56, No. 1, 67–79 (2008)

Phylogenetic Analysis Using Torus Self-Organizing Map

Shinichi Horata[1], Toshimichi Ikemura[2] and Tetsuyuki Yukawa[3]

[1]Hayama Information and Network Center, The Graduate University for Advanced Studies
[2]Nagahama Institute of Bio-science and Technology
[3]Hayama Center for Advanced Studies, The Graduate University for Advanced Studies

To address the problem of clarifying interspecies difference of genome sequences, Self-Organizing Map (SOM) was used as a classification method concerning species and phylotype families. In order to clarify relation of clusters with species and phylotype families, we analyzed the position of each cluster on a SOM. We employed the torus map algorithm, which provided independence of the position on the map, and compared with the results obtained by the plane map algorithm. We performed SOM analysis for genome sequences of influenza virus (11585 2-kb genomic fragments). The numerical results suggested that the torus map could make clear the relation between clusters and characterize features of the phylotype families after learning.

Key words: SOM, Self-Organizing Map, bioinformatics, genome informatics, oligonucleotide frequency.