メタゲノム解析による微生物群集構造の解明への一括学習型自己組織化マップ（BLSOM）の活用

阿部　貴志[1]*・中尾　亮[2]・杉本　千尋[2]

次世代シーケンサーに代表されるゲノム解読技術のハイスループット化に伴い，「メタゲノム解析」と呼ばれる新実験分野を生み，全地球レベルでの生物生態系の把握を目標にした大規模解析が行われている．「メタゲノム解析」とは，環境中に生息する生物群集に由来する，多種類のゲノムの混合物を対象にした大規模ゲノム配列解読である．近年，新型シーケンサーの一つであるロシュ社製GS FLXを用いて，平均400塩基のメタゲノム配列の大量解読を行う場合が多い．

大量のメタゲノム配列を用いて，生物生態系を正確に把握し，環境保全や修復に係わる知識を生産・活用することは，人類にとって重要な課題の一つといえる．しかし，メタゲノム配列は，新規性の高い配列が多く含まれることから既存の類似配列が少なく「どの系統の生物種に由来するのか」を配列相同性検索のような従来法で適確に推定することは困難であり，生物種情報や機能情報などの情報が付与されない素データのまま公開されている．効率的に知識発見を行うための情報処理システムの確立が重要である．

そこで，既知の全ゲノム配列を対象に，各生物を特徴付けるゲノムサイン（genome signature）を把握するために開発した連続塩基組成に着目した一括学習型自己組織化マップ（Batch-Learning Self-Organizing Map, BLSOM）を応用し，環境中の未知難培養性微生物ゲノム配列断片の生物系統推定法を開発した[1–3]．現在，世界的に普及し始めており，連続塩基組成や自己組織化マップに着目した生物系統推定法が，他グループからも提案されている[4–7]．

本稿では，開発したBLSOMを用いたメタゲノム配列に対する生物系統推定法を活用した研究について紹介する．

連続塩基組成に着目した一括学習型自己組織化マップ（BLSOM）によるゲノム配列解析

ゲノム配列は，タンパク質をコード化する情報以外にも，多様な情報を含んでいる．すべてのゲノムはACGTの4種の文字で書かれているので，断片的な配列（たとえば10,000塩基）だけが与えられたのでは，どの生物の配列なのかを知ることは不可能に思える．GC含量は個々のゲノムの基本的特徴として使用されてきたが，多様なゲノムを区別するためのパラメータとしては単純すぎる．Karlinらは，2連塩基（dinucleotide）頻度が広範囲な生物種間で明瞭に異なり，ゲノムサイン（Genome Signature）が存在することを示した[8]．同一のGC含量を持つ生物種でも，2連塩基の同一な出現頻度特性を持つ生物種は少ないはずであり，連続塩基が3連や4連と長くなるにつれ，同一の頻度特性を持つ生物種の可能性は極端に小さくなる．この連続塩基頻度解析を，国際塩基配列データベース（INSD）に収録されているゲノム配列の全体を対象にして，大規模な解析を行うことで，新規な視点での知識発見が可能となった．

筆者らは，生物種固有のゲノムサインの全体像を把握することを目的に，大量かつ多次元ベクトルデータを2次元マップ上でクラスタリングと視覚化できるコホネンらが開発した教師なし学習アルゴリズム「自己組織化マップ（Self-Organizing Map，SOM）」に着目した[9]．コホネンSOMの長所を生かしながら，再現性のある分類結果を得るアルゴリズムとして「一括学習型自己組織化マップ（Batch-Learning Self-Organizing Map, BLSOM）」を開発し，ゲノム配列解析に適用してきた[1,2]．

BLSOMは，各ゲノム配列を断片化（たとえば，1000塩基ごと）し，そのゲノム配列断片中の連続塩基（たとえば，3連や4連塩基）出現頻度を入力データとし，各ゲノム配列断片間の連続塩基出現頻度の類似度のみで2次元の格子状のマップ上でクラスタリングを行う．クラスタリングされた各ゲノム配列断片の生物種情報を見ると，生物種に関する情報を一切計算機に与えることなく，ゲノム配列断片を高精度に生物系統別に分離（自己組織化）されていることが判る．たとえば，図1Aは，現時点で10 kb以上の配列断片がINSDに収録されている原核生物の約5600種，既知真核生物の412種，既知ウイルス約30,000種，ミトコンドリア4479種，葉緑体225種について，塩基配列を5000塩基（5 kb）ごとに断片化したゲノム配列断片1120万件（56ギガ塩基）を対象に，縮退4連塩基頻度に基づくBLSOMを地球シミュレータにて解析した結果である（ここでは，Kingdom-BLSOMとする）．ここでは，相補的な連続塩基（たとえばAAAA

*著者紹介　[1]新潟大学大学院自然科学研究科（准教授）　E-mail: takaabe@ie.niigata-u.ac.jp
[2]北海道大学人獣共通感染症リサーチセンター

とTTTT）を同一のものとみなすことを，「縮退」と定義している．高等生物の巨大ゲノムの断片化配列をそのまま解析に加えると，それらの配列の寄与が大きくなりすぎるため，ゲノム配列長が200 Mb以上の高等生物種については，200 Mb分を計算機でランダムに抽出し解析データとした．このことで，原核と真核生物の配列の総量をほぼ等量にしている．BLSOMの学習後に，格子点が同一の生物系統の配列のみの場合にはその系統を示す色で，複数の系統の配列が混在する場合には黒色で示した．大半の格子点が生物系統ごとに色付けされており，真核と原核生物については95%と高精度で分離されていた．生物系統を反映してゲノム配列断片がクラスタを形成（自己組織化）していることが明らかである．

また，ウイルスゲノムのクラスタも形成されており，固有のゲノムサインを持つことが示唆された．そこで，全ウイルス44,000株の配列を対象に，1 kbと500 bに断片化し，4連塩基でのBLSOM解析を行った．ウイルスのOrder（目）ならびにFamily（科）の系統レベルでの分類結果を図1Bに示す．断片化サイズ1 kbと500 bのFamilyレベルでの分離は，99%と98%以上と非常に高精度に分類されていた．BLSOMを用いることで広範なウイルス間での比較ゲノム解析が可能である[10]．

BLSOMは，教師なしのアルゴリズムであり，ゲノム配列断片中の連続塩基出現頻度のみに着目することで，予備知識なしにゲノム配列に潜む生物種固有の特徴（ゲノムサイン）をクラスタリング･可視化することができ，着目した生物が持つシグナルやモチーフを含む特徴配列や水平伝播遺伝子候補の探索にも活用できる[11,12]．

BLSOMを用いたメタゲノム配列に対する系統推定法

大量のメタゲノム配列情報から環境微生物群集を迅速にモニターし，科学的・産業的に興味深い有用遺伝子を体系的に発掘することがメタゲノム解析の目指すべき課題である．予備知識なしにゲノム配列断片を高精度に生物系統別に分離できるBLSOMを応用したメタゲノム配列に対する生物系統推定法について紹介する[3]．原核生物を対象にした系統推定のワークフローを図2に示す．

自然環境試料のメタゲノム解析では，新規性の高い原核生物だけではなく真核生物のゲノムDNAが混入している可能性が考えられ，メタゲノム配列に対して系統推定するためには，現時点で塩基配列が解読されているすべての生物種に対し，連続塩基出現頻度の特徴をBLSOMで予め把握しておく必要がある．そこで，3つの異なる系統レベルに対するBLSOMとして，図1で示したウイルスやオルガネラを含むすべての生物種のKingdom-BLSOM，全原核生物を用いたPhylum（門）レベルでのBLSOM（Prokaryote-BLSOM），各PhylumでのGenus（属）レベルでのBLSOM（Genus-BLSOM）を断片化サイズ5 kb，縮退4連塩基にて作成した．

300塩基以上のメタゲノム配列を対象に，Kingdom-BLSOMへマッピングを行ない，マップされた格子点とその近傍に分類されていた既知生物種情報を用いて，生物ドメイン（真核生物，原核生物，ウイルス，ミトコンドリア，葉緑体）を推定する．次に，原核生物と推定されたメタゲノム配列を用いて，Prokaryote-BLSOMへマッピングを行ない，原核生物のPhylumレベルを推定する．さらに，Phylumを推定できたメタゲノム配列を用いて，各PhylumでのGenus-BLSOMへマッピングを行うことで，属や種レベルでの系統推定が可能となる．このように真核生物・原核生物などの生物ドメイン，原核生物の系統群，属種レベルと階層的に生物系統を絞り込んでゆくことで，より詳細な系統推定が可能である．さらに，各BLSOMへマッピングを行った際に，生物系統を推定できない場合もあるが，どの系統レベルかを知ることができるため，新規性の高いメタゲノム配列の効率的な検出ができる．

ここで紹介したBLSOMを用いた系統推定のワークフローについては，メタゲノム配列を投入するだけで，上述のBLSOMマップを用いた階層的な系統推定を自動的に実行できるソフトウェアPEMS（phylogenetic estimation of metagenomic sequence using BLSOM）として公開している（http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft）．

また，マッピングを行うBLSOMマップを換えることで，真核生物やウイルスなどの他の生物系統についても推定可能である．特に，ウイルスゲノムにはrDNAが存在しないので，系統推定法として広く普及しているrDNA配列を用いた系統推定は困難であるが，ウイルス由来のメタゲノム配列の選別と系統推定に活用できる．

本手法は，新規性の高い遺伝子類の配列で信頼性のある系統樹を作成するのに必須となる，広範囲の生物系統をカバーするオーソログス配列セットを必要としないアライメントフリーな手法であり，配列相同性検索と異なった原理に基づく手法である．また，異なる試料間での生物群集の多様性比較もできる．さらに，BLSOMを用いたメタゲノム配列に対するゲノム別の再構築法の開発[13]も行なっており，系統推定法と組み合わせることにより，既知微生物類とは類似性を示さない真に新規性の高い微生物ゲノムも検出できる．さまざまな環境が持つ生物群集システムの比較解析や特定の環境中に生息す

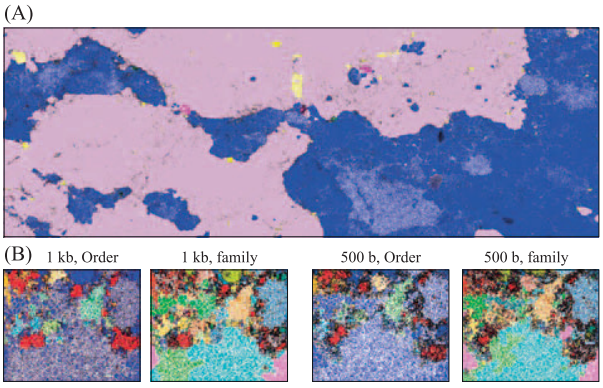

図1．BLSOM分類結果．（A）ウイルスやオルガネラを含む全ての生物種を対象にした縮退4連塩基でのBLSOM．ここで，色と系統の対応は以下のとおりである．■：原核生物，■：真核生物，■：ウイルス，■：ミトコンドリア，■：葉緑体．（B）全ウイルスを対象にした4連塩基でのBLSOM．色と系統の対応は以下のとおりである．Order: ■：*Caudovirales*, ■：*Herpesvirales*, ■：*Mononegavirales*, ■：*Nidovirales*, ■：*Nidovirales*, ■：*Orderunclassified*. Family: ■：*Coronaviridae*, ■：*Siphoviridae*, ■：*Hepadnaviridae*, ■：*Flaviviridae*, ■：*Poxviridae*, ■：*Retroviridae*, ■：*Orthomyxoviridae*.

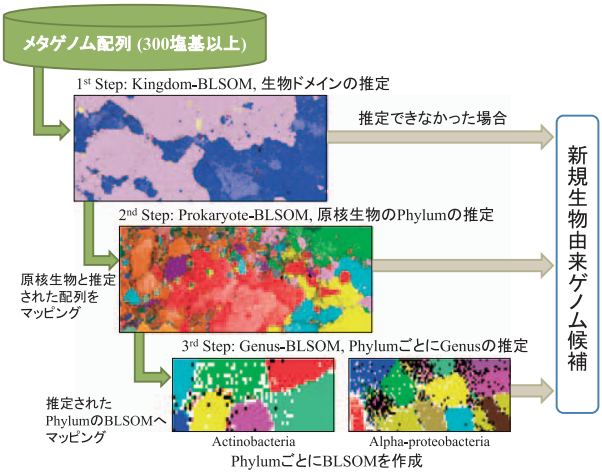

図2．BLSOMを用いた生物系統推定法のワークフロー

る各生物が持つ代謝経路の概要の推定を行うことで，難培養性微生物のゲノム資源の活用においても基礎的な有用情報を提供できる．

メタゲノム解析によるマダニ保有微生物叢の探索

マダニはさまざまな病原性ウイルス，細菌，原虫を保有することが知られており，ヒトや動物を吸血することでその伝播に関わる．特に1991年以降に報告されたマダニ媒介性*Rickettsia*属細菌によるヒトの新興感染症は12例にも上がり，さらに2011年には中国でヒトへの病

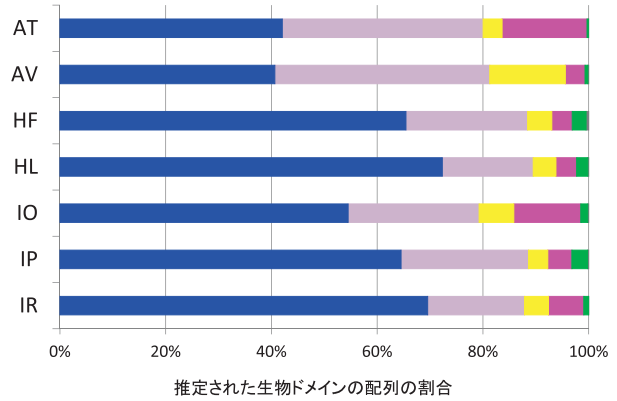

図3．推定された生物ドメインの分布．色と生物ドメインの対応は，図1（A）を参照．

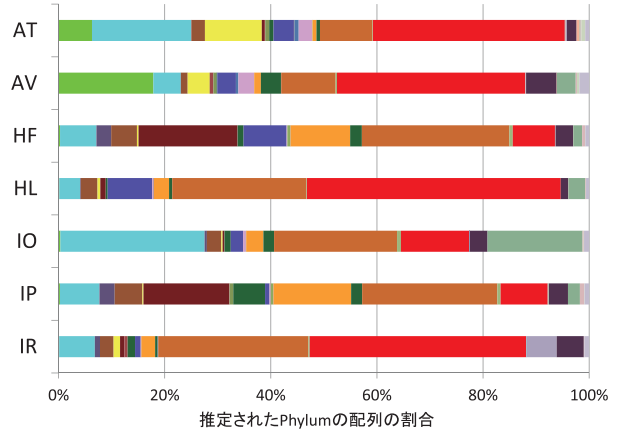

図4．推定された原核生物Phylumの分布．代表的なPhylumと色の対応は以下のとおりである．■：Actinobacteria, ■：Alpha-proteobacteria, ■：Beta-proteobacteria, ■：Chlamydia, ■：Cyanobacteria, ■：Epsilon-proteobacteria, ■：Firmicutes, ■：Gamma-proteobacteria, ■：Spirochaetes, ■：Tenericutes.

原性がきわめて高い新規ブニヤウイルスがマダニから検出されるなど，未知の病原体が潜んでいる可能性がある．マダニ媒介性病原体のほとんどは，ヒトや動物の病原体として認識される以前にマダニから検出されている．マダニが保有する微生物叢を網羅的に解析することは，未知の病原体の存在を明らかにできるとともに，新興のマダニ媒介性感染症の先回り対策として有効な手段と考えられる．そこで，次世代シーケンサーを用いたメタゲノム解析を行ない，哺乳動物には病原性を持たない共生・寄生体に加えて，未知の病原体，新興感染症を引き起こす可能性のある微生物を含めマダニ保有微生物叢の解明を試みた．

野外採集（北海道，宮崎県，オランダ・ユトレヒト）

表1．メタゲノム解析により取得した配列数とBLSOMでの系統推定結果

	AT	AV	HF	HL	IO	IP	IR
配列数	21,639	69,840	21,213	24,249	73,211	70,925	39,810
平均長	149.0	277.5	185.0	187.5	267.2	257.6	149.9
300塩基以上の配列数	2906	30,361	4502	5186	32,766	32,828	5438
原核生物と推定された配列数（%）	1222 (42.1)	12,340 (40.6)	2946 (65.4)	3750 (72.3)	17,859 (54.4)	21,179 (64.5)	3782 (69.5)
Phylumが推定された配列数（%）	1213 (99.3)	12,113 (98.2)	2925 (99.3)	3728 (99.4)	17,686 (99.0)	20,995 (99.1)	3748 (99.1)
Genusが推定された配列数（%）	1119 (92.3)	10,487 (86.6)	2605 (89.1)	3319 (89.0)	15,820 (89.4)	18,697 (89.1)	3398 (90.7)

ならびに実験室継代コロニー（ガンビア由来）のマダニ7種（*Amblyomma testudinarium* (AT), *Amblyomma variegatum* (AV), *Haemaphysalis formosensis* (HF), *Haemaphysalis longicornis* (HL), *Ixodes ovatus* (IO), *Ixodes persulcatus* (IP), *Ixodes ricinus* (IR)）について，遠心分離・フィルター処理により細菌画分を濃縮・精製し，得られたゲノムDNAを材料にRoche/454 GS FLXで塩基配列の解読を行った．300塩基以上のメタゲノム配列を用いて前述のBLSOMによる系統推定を行った．取得された配列件数と系統推定できた配列の件数を表1に示す．Kingdom-BLSOMにマッピングを行い，生物ドメイン（原核生物，真核生物，ウイルス，ミトコンドリア，葉緑体）を推定した結果，99.5%以上がそのいずれかに推定された（図3）．大半が原核生物に推定されていたが，真核生物に推定された配列のほとんどはマダニに由来すると予想された．原核生物由来と推定された配列をProkaryote-BLSOMへのマッピングによって，原核生物のPhylumレベルを推定した結果（図4），それぞれのマダニ種で異なる構成比が得られた．全体の傾向として，FirmicutesとGamma-proteobacteriaに属する細菌由来の配列が全体の約半数を占めた．*Rickettsia*などが属するAlpha-proteobacteriaはほとんどすべてのマダニ種でみられた．一方で，これまでマダニで報告のなかったChlamydiaeに属する配列の割合が多いマダニ種が見られた．Genus-BLSOMにより，原核生物のGenusレベルの推定を行ったところ，これまで報告のある，マダニ媒介性病原細菌（*Anaplasma*属，*Bartonella*属，*Borrelia*属，*Ehrlichia*属，*Francisella*属，*Rickettsia*属）ならびにマダニ共生細菌（*Coxiella*属，*Rickettsiella*属，*Wolbachia*属）が検出された．さらに，潜在的病原体の候補として，*Chlamydia*・*Chlamydophila*属由来が一部のマダニ種にみられた．これらについては特異的プライマーを用いて16S rDNAの配列決定も行い，その存在を確認できた．

メタゲノム解析による微生物叢の網羅的解明は，環境試料中の有用微生物や遺伝子の探索のみならず，自然界に存在する未知の病原体あるいは潜在的に病原性を持つ新規性の高い微生物を見つけ出すことが可能であり，新規感染症の原因微生物探索を含む医学・医療分野にも活かすことができる．

おわりに

大規模メタゲノム解析において，迅速かつ体系的に生物系統やタンパク質機能情報を付与するためのバイオインフォマティクスは重要な役割を占める．本稿では，BLSOMを用いた生物系統推定法について紹介したが，BLSOMをタンパク質アミノ酸配列解析に適用し，配列相同性に依存しないタンパク質機能推定法の開発を行なっており，メタゲノム配列からの科学的・産業的に有用なタンパク質遺伝子の発掘に応用できる[14]．微生物生態系を把握し，微生物が持つ代謝システムの体系的な把握に向け，環境情報や多様な実験データと有機的に結びつけていくことも重要となってくるだろう．

文　献

1) Kanaya, S. *et al.*: *Gene*, **276**, 89 (2001).
2) Abe, T. *et al.*: *Genome Res.*, **13**, 693 (2003).
3) Abe, T. *et al.*: *DNA Res.*, **12**, 281 (2005).
4) Teeling, H. *et al.*: *BMC Bioinformatics*, **5**, 163 (2004).
5) McHardy, A. C. *et al.*: *Nat. Methods*, **4**, 63 (2007).
6) Chan, C. K. *et al.*: *BMC Bioinformatics*, **9**, 215 (2008).
7) Dick, G. J. *et al.*: *Genome Biol.*, **10**, R85 (2009).
8) Karlin, S. *et al.*: *J. Bacteriol.*, **179**, 3899 (1997).
9) Kononen, T.: *Proc. IEEE*, **78**, 1464 (1990).
10) Iwasaki, Y. *et al.*: *DNA Res.*, **18**, 125 (2011).
11) Kosaka, T. *et al.*: *Genome Res.*, **18**, 442 (2008).
12) Yasui, K. *et al.*: *Biosci. Biotechnol. Biochem.*, **73**, 1422 (2009).
13) Uehara, H. *et al.*: *Genes Genet. Syst.*, **86**, 53 (2011).
14) Abe, T. *et al.*: *DNA Res.*, **16**, 287 (2009).