超大型スーパーコンピュータで可能になるゲノム情報に基づく生物の俯瞰的把握と環境ゲノム資源活用のための情報学的手法確立

阿部貴志[1], 金谷重彦[2], 池村淑道[1]

[1].長浜バイオ大学, [2].奈良先端科学大学院大学
E-mail: takaabe@nagahama-i-bio.ac.jp

概要：我々は、ゲノム配列のオリゴヌクレオチド頻度のみで断片配列の生物種による高精度な分類が可能である一括学習型自己組織化マップ（BL-SOM）を用いて、環境微生物ゲノム断片配列の系統分類や新規性の推定、ならびに未知微生物由来配列を効率的に特定する手法の開発を行った。その手法を用いて、メタゲノム解析によって得られた土壌・海洋由来の大量な DNA 断片配列を対象にして、培養が困難な微生物の混合試料に由来するゲノム断片配列の系統推定、異なる環境間での微生物群集の比較ゲノム解析における有用性を明らかにした。

1　はじめに

ゲノムプロジェクトの進展に伴い、現在では600を超える生物種の完全なゲノム塩基配列が明らかにされ、これらのゲノム情報を基に生物種の個性を塩基配列レベルで把握することが可能となった。多様な環境に生息する多種類の生物種を対象にした、以下のような新規視点でのゲノム解析も注目を受けている。極限環境を代表例とする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る。これらは新規な遺伝子類を豊富に保有する可能性があり、注目を集めている。この難培養性微生物を解析する目的で、環境中で生育する生物群を含有する試料から培養せずにゲノムDNAの混合物を直接に抽出し、配列決定を行う技術(メタゲノム解析)が開発され、世界的に普及してきた。しかしながら、得られたゲノム断片配列の集合のみでは、各配列が由来する生物の種類、系統群、さらにそれらの新規性を推定することは困難である。我々が開発した、オリゴヌクレオチド頻度のみでゲノム断片配列の生物種による高精度な分類(自己組織化)が可能な、一括学習型自己組織化マップ（BL-SOM)を用いて［1,2]、環境微生物集団のゲノムの多様性や新規性を推定するための系統分類法、ならびに新規性の高い未知微生物ゲノムを効率的に探索するための新手法を開発した[3,4]。

2　方法

コホネンが開発した自己組織化マップ（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。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、ゲノム塩基配列の解析には殆ど用いられずにきた。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。我々は、従来型のコホネン SOM の長所を生かしながら、再現性のある分類結果を得る形式にアルゴリズムを変更するために、「一括学習型の自己組織化マップ法(BL-SOM)」を開発してきた。大量データに対する大規模な並列処理が可能となり、次世代スーパーコンピュータによる大量データ解析に適したアルゴリズムとなった。

3 超大型スーパーコンピュータで可能になるゲノム情報解析

環境試料のメタゲノム解析で得られた、生物系統が未知の難培養性微生物に由来するゲノム断片配列の系統を推定するためには、現時点でデータベースに収録されている既知生物種の全 DNA 配列を、予め BL-SOM で分離しておく必要がある。自然環境試料を対象にしたメタゲノム解析では、原核生物だけでなく真核生物のゲノム DNA が混入している可能性が高い。さらには、メタゲノム解析が殺菌した臨床試料へも適用可能なことから、新規感染症の原因となる未知病原微生物の探索にも利用可能である。このような医薬学分野での混合ゲノム試料を対象した場合は、ヒトのみならず広範囲の真核生物由来の DNA の混入が想定される。さらには、ウイルスやミトコンドリアやクロロプラストやプラスミド等をも含む、既知の全塩基配列を対象にした大規模な BL-SOM を作成しておくことが望ましい。図 1a では、現時点で 10kb 以上の断片配列がデータベースに収録されている約 1,512 種の原核生物に加えて、約 40 種の真核生物、約 1,600 のウイルス、約 600 のオルガネラ(ミトコンドリアや葉緑体)の塩基配列の全体を対象に、塩基配列を 5,000 塩基ごとに断片化(総入力データ数 約 40 万件)し、縮退させた 4 連塩基の出現頻度(136 次元のベクトルデータ)について地球シミュレータを用いて解析した。なお、現在は、約 3,000 種の微生物、約 110 種の真核生物、約 36,000 のウイルス、約 1,800 のオルガネラの計 360 万件を対象にした同様の解析を行っている。縮退とは、2 本鎖 DNA 配列の相補性の影響を除去するために、相補的なオリゴヌクレオチド（例えば、AAAA と TTTT）を同一のものとみなすことを意味している。但し、地球シミュレータの性能上の限界から、高

等動植物の巨大ゲノムの全配列を加えると、データ数が大量すぎて、現実的な時間内での計算が不可能である。図 1a で紹介する解析の場合、主対象が難培養性の微生物類の新規ゲノム配列の系統推定であるので、生物種既知の真核生物について、ヒトのような大型ゲノムの場合は 200Mb 分（ゲノム全体の 1/10 以下）の配列をランダムに選択して解析に使用している。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が、真核と原核生物については 95%レベルの高精度で分離されている。オルガネラとウイルス相互や、これらと核ゲノムとの分離も 80%レベルと高い。また、約 3,000 の既知原核生物に関して、28 の系統群への分離を調べると、85%レベルで各系統群を反映した領域に分離していた。100%の分離には至っていないが、その主原因は異なったゲノム間での遺伝子類の水平伝播に起因すると考えている。図 1b では、バーミューダ近海から汲み上げた大量な海水から得た多数の生物種の混合ゲノム試料に由来する、大量なゲノム配列の断片（21 万件）を図 1a の BL-SOM へマップしている。90%程度が原核生物の領域にマップされていた。

人類が現時点で知っている、ウイルスやミトコンドリアを含む全ゲノム配列を一枚の BL-SOM 上で分離（自己組織化）し公開することは、医薬学を含む広いライフサイエンス分野のみならず、関連の産業分野に対しても、世界的に類例にない新規で大規模な基盤ゲノム情報の提供となる。各研究者が新たに解読した配列類を、PC レベルの計算機で、この公開した大規模 BL-SOM へマップすることで、着目配列類の系統推定が可能になる。

4　まとめ

現時点でゲノム配列解読が進行している真核生物は 500 種を超えているが、地球シミュレータの性能上の限界から、図 1a の例では、40 種の真核生物を解析しているに過ぎない。メタゲノム解析における有用性が微生物に限定されてしまう。次世代スーパーコンピュータを用いて、その時点で配列が既知の全ゲノム配列を一枚の BL-SOM 上で分離し公開すれば、各実験グループが取得した環境由来のゲノム断片配列類について、各自が PC レベルの計算機でこの BL-SOM 上にマップすることで、生物系統が推定できる。環境由来の未開拓ゲノム資源活用のために、世界の追従を許さない大規模な知識情報を提供でき、ライフサイエンスの基礎分野だけでなく、医薬学を含む産業分野への大きな貢献となる。

SOM は教師なしのアルゴリズムであり、断片配列の大半が連続塩基の出現頻度の類似度のみで、予備知識なしに分離し、ゲノム配列に潜む生物種の特徴を明らかにできる。単一のゲノムに着目すると、ゲノムの機能領域ごとに分離する傾向を示し、機能領域の配列に潜む特徴を明らかにできる。

生物のゲノム配列に内在する生物種ごとの特徴を把握することは、生物の多様性やゲノム進化に関する基礎知識を得る上でも重要である。我々が開発した方法は、従来の系統推定法と異なり、オルソログ配列のセットや配列間のアラインメントが必要でない。連続塩基の出現頻度のみで系統推定が可能なので、新規性の高い未知な生物種の配列類の系統推定には最適な方法である。現在 100 種類を超える環境試料を対象にした大規模メタゲノム解析が実施されており、本手法を用いることによって、環境中に生息する微生物集団の実体の解明や環境間での効率的で正解な比較が可能となる。

超高速で超並列な大規模計算を可能にする次世代スーパーコンピュータによる大量ゲノムデータを対象にした BL-SOM 解析とその公開は、広範囲のライフサイエンス分野と産業分野の研究開発を支援するための、先端的で応用範囲の広い、世界で類例のない基盤ゲノム情報を提供できる。

参考文献

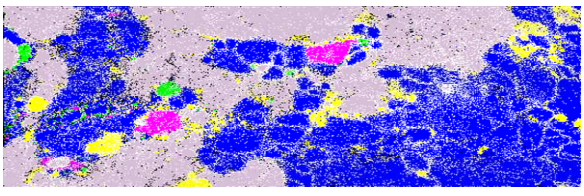
[1] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura.「Analysis of codon usage diversity for bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome.」, *Gene*, 276, 89-99, 2001.

[2] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, T. Ikemura,「Informatics for unvailing hidden genome signature.」, *Genome Res.*, 13, 693-702, 2003.

[3] T. Abe, H. Sugawara, M. Kinouch, S .Kanaya, T. Ikemura,「Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples.」, *DNA Res.*, **12**, 281-290, 2005.

[4] T. Abe, H. Sugawara, S .Kanaya, T. Ikemura,「Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator」, Journal of the Earth Simulator, 6, 17-23, 2006.

図1(a) 原核生物と真核生物、ミトコンドリア、葉緑体、ウィルスのBL－SOM



(b)サルガッソ海の混合ゲノムに由来する大量断片配列のマップ結果