全ゲノム情報に基づいた病原微生物と常在菌の多様性と病原遺伝子に関する情報学的研究

●池村 淑道　◆阿部 貴志
長浜バイオ大学・バイオサイエンス学部

＜研究の目的と進め方＞

生体内を含む環境中に生息する微生物類の多くは培養が困難であり、実験室系のアプローチが困難であったため、未知・未開拓に残されてきたゲノム群といえる。難培養性微生物に関する研究が困難であったため、感染や病原性メカニズムの解明、病原菌と常在菌の相互作用を研究する上での基盤ゲノム情報が整備されずにきた。最近、微生物類を培養することなく混合ゲノム DNA として回収し、大量の断片 DNA 配列を解読する方法論「メタゲノム解析」が開発されている。新規性の高い広範囲のゲノムが解析でき、滅菌消毒を行った試料についても微生物類の検出が可能となる。しかしながら、混合ゲノム試料から得られた大量の断片配列の集合のみでは、各断片配列が由来した生物の系統、特に新規性の高い配列の生物系統を推定することは困難であった。我々のグループは、SOM(自己組織化地図) 法に改良を加えた BLSOM 法を開発し、この弱点を克服する新規な情報学的手法を確立し、難培養性微生物由来配列の系統推定に応用してきた。このゲノム解析技術を発展させて、常在菌の多様性や量比の正確な把握、新規性の高いゲノムに由来する配列の特定、ウイルスを含む病原微生物の探索や特定等を可能にする実践的な技術として確立し、急速に増加を続けている配列情報の全体を対象にした BLSOM を継続的に更新し、感染や病原性メカニズムの解明のための基盤ゲノム情報を提供する。

生体内を含む環境中に生息する微生物類を対象にしたメタゲノム解析の主目的は、生命科学的ならびに産業的に有用なタンパク質遺伝子を新規性の高いゲノムから発掘することにある。このメタゲノム解析の配列解読の場合でも、単一ゲノムの配列が解読された場合でも、新たに得られたタンパク質遺伝子の機能については、機能既知なタンパク質のアミノ酸配列との配列相同性を根拠に推定するのが一般的である。しかしながら、タンパク質の機能については、機能部品類の 3 次元上での立体配置が重要であり、同一ないしは類似の機能を持つタンパク質間でも、アミノ酸の 1 次元配列上での全域に渡る相同性を見付けられない例が多い。現在では、機能が配列相同性検索で推定出来ない例が多数集積しており、新規性の高いゲノムを対象にするメタゲノム解析では、その割合が特に強い。既に 500 万件を超える機能が未知の遺伝子候補が、利用価値の低いままにデータベースに集積している。この状況を解決する視点から、タンパク 3000 プロジェクトで代表されるように、X 線結晶構造解析や NMR 法でタンパク質の 3 次元構造を決定し、機能既知なタンパク質との高次構造上の類似性で機能を推定する大規模プロジェクトが進行している。しかしながら、費用や労力ならびに技術上の諸問題から、今後ますます蓄積する機能未知な多数のタンパク質類の機能推定には不十分と考えられる。配列相同性検索を補完する、異なった原理に基づくタンパク質の情報学的な機能推定法の確立が急務と言える。

約 10 年前に、タンパク質の 2 連アミノ酸頻度を SOM 解析した欧米のグループの研究により、高次構造や機能による分離が起きることが報告されている。タンパク質の機能推定に有用と考えられるが、ゲノム配列解読前の研究であり、機能未知のタンパク質がほとんど知られていなかったことや、長時間の計算が必要なこと、得られる最終マップが初期条件や入力データの順番に依存すること等で、タンパク質の配列解析には殆ど用いられずにきた。本研究開発では、我々のグループが開発した入力データ順に依存しない改良型の SOM である BLSOM を用いて、タンパク質の機能推定法としての可能性を検証する。予備的な小規模な解析では、ダイやトリペプチドに着目した BLSOM により、タンパク質は機能による分離 (自己組織化) の傾向を示した。配列相同性検索とは異なる原理に基づくタンパク質の機能推定法として確立を目指す。

＜ 2007 年度の研究の当初計画＞

難培養性の常在ならびに病原微生物類の多様性と系統推定のための、既知の全ゲノム配列を対象にしたBLSOMの更新とタンパク質機能推定用のBLSOM開発。

感染症の病原性や感染のメカニズムの解明には、ヒトのみならず広範囲の生物試料由来の混合ゲノム DNA 解析が想定される。真核と原核生物ゲノムのみならず、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の既知の全塩基配列を対象にした大規模な BLSOM を作成し、更新することが重要である。この目的の大規模 BLSOM について、4 ～ 5 連塩基頻度解析の高速化を計り、高機能な表示システムを付加して、広く利用可能なシステムとして完成させる。

微生物ゲノム由来の大量なタンパク質配列を対象に、地球シミュレータを用いて多様な解析条件での大規模 BLSOM 解析をおこない、タンパク質の機能推定法として最適な解析条件を決定する。得られた解析条件で、病原微生物と常在菌ゲノムに由来する多数の機能が未知遺伝子の機能推定を試み、メタゲノム解析で得られた遺伝子の候補類に関しては、由来するゲノムの BLSOM による系統推定を行う。特に、BLSOM 上で、病原性との関係性が示唆されている既知の遺伝子類と一緒に分離 (自己組織化) する機能未知遺伝子類に着目して、推定された生物系統と合わせたカタログ化を行う。

＜ 2007 年度の成果＞

現時点で 10kb 以上の断片ゲノム配列がデータベースに存在する 2000 種を超える既知原核生物に加えて、約 1100 種類のウイルス、配列解析の進んでいる約 50 種類の真核生物、約 700 種類のオルガネラ配列の全体を 5kb に断片化し、地球シミュレータを用いて、4 連塩基頻度に関する BLSOM 解析を行った。真核と原核生物については 96% の高精度で分離 (自己組織化) していた。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。2000 種を超える既知原核生物種に関して、25 の系統群への分離の度合いを調べると、85% レベルで正しい系統群を反映して分離していた。体内環境や共生生物系を含む環境由来の大半のゲノム断片の系統推定を可能にできた。原核生物で系統群を間違えて分離した 15% の配列上には、水平伝播をした外来性的な配列が多く見出され、病原性と関係する遺伝子群が濃縮さ

れている可能性が高い。これらのカタログ化を行った。

タンパク質遺伝子のコドン（遺伝暗号）使用頻度パターンも生物種ごとに明瞭な特徴を持っており、特に微生物ではその特徴が顕著である。水平伝播をした外来性的な遺伝子のコドン（遺伝暗号）使用パターンは、acceptor 側ゲノムの特徴よりも donor 側ゲノムの特徴を持つことが知られている。コドン使用頻度に関する BLSOM 解析を行なうことで、外来性遺伝子ならびにその donor ゲノムの候補の推定が可能になった。我が国の複数の微生物ゲノム解析グループとの共同研究が進展しており、一部は既に論文発表を行なった（*Genome Research*, in press）。

ヒトやマウスの腸内試料を対象にした大規模メタゲノム解析が進行しており、東大の服部グループのヒト腸内試料の約 730Mb 分の配列が国際 DNA データベースに登録されたことで、総量が 1Gb 分に達している。これらの登録されている大量断片ゲノム配列についての BLSOM 解析を行なっているが、ヒトやマウスの腸内試料で見出された生物多様性は、サルガッソ海を代表とする海洋試料やミネソタの土壌試料に比べれば遥かにその多様性は低かったが、個体差とその特徴が特定できた。大量配列を対象にしたアッセンブル法により優先種のゲノムの概要が再構成可能なことが知られているが、BLSOM の場合には比較的少量しか存在しない生物種由来のゲノム配列についてもゲノム別での再集合を可能にしている。

機能未知のタンパク質類の機能推定法を確立する目的で、既に機能が特定された 2853 種類の COG カテゴリーへ分類がなされている、完全ゲノム配列が解読された微生物ゲノム由来の約 11 万個のタンパク質類のアミノ酸配列をテストデータとして用いた。2 ～ 4 連アミノ酸頻度を対象に、地球シミュレータを用いて大規模 BLSOM を作成し、精度高く COG への分類を実現する計算条件を求めた。タンパク質の BLSOM については機能と生物系統の両方を反映する分離が混在したが、アミノ酸を物理化学的な性質の類似度でグループ化することで、機能を反映した分離の程度が増大することを見出した。本年度の解析では、20 種類のアミノ酸での 2 連アミノ酸頻度 (400 次元のベクトルデータ)、11 カテゴリーへグループ化した後の 3 連アミノ酸頻度 (1331 次元のベクトルデータ)、6 カテゴリーへグループ化した後の 4 連アミノ酸頻度 (1296 次元のベクトルデータ) を主に試みた。これらの 3 種類の BLSOM 解析条件の内で、11 カテゴリーへグループ化した 3 連アミノ酸頻度の BLSOM が最も機能を反映して高い分離能を与えていた。タンパク質は多様な長さがあるが、200 アミノ酸の Window を設け、50 アミノ酸の Step で移動することで、大型タンパク質も解析を可能にした。現時点で蓄積している大量の機能未知のタンパク質類について、機能機知のタンパク質類と混合して大規模 BLSOM を作成し、機能未知と既知タンパク質がマップ上でアソシエートすることを指標に機能推定を試みている（論文作成中）。上記の 3 種類の BLSOM 解析条件で同じ機能が推定できたタンパク質から公開をする予定である。特に、病原性と関係することの知られている機能既知の遺伝子とアソシエートする機能未知な遺伝子類の大規模な探索を行なっている。

＜国内外での成果の位置づけ＞

BLSOM を用いた環境微生物由来のゲノム断片配列の生物系統の推定法については、発明の名称「塩基配列の分類システムおよびオリゴヌクレオチド出現頻度の解析システム」として特許が認定されており、独創性の高い情報学的手法を確立できたことから、我々のグループが中心になって解析を進めている。既に我が国の 7 箇所の環境ゲノム解析の実験グループからの解析の依頼を受けており、共同研究の一部の成果の発表を終えている（*Genome Research* in press; *Nature Biotechnology* 2005; *Canadian Journal of Microbiology* 2005）。

機能未知のタンパク質の BLSOM を用いた機能推定法も、我々のグループが独自に開発した方法であり、特徴ある解析手法の導入と言える。オリゴペプチド頻度はオリゴヌクレオチド頻度と同様に高次元なベクトルデータである。機能未知のタンパク質についての機能推定であれ、生物種未知の微生物由来のゲノム断片配列についての由来生物の系統推定であれ、現時点で公的データベースに収録された全ての既知生物種や全ての既知タンパク質の特徴抽出を、予め行なっておく必要がある。従って、大量な高次元ベクトルデータの BLSOM 解析が必要となるので、地球シミュレータレベルの高機能スーパーコンピュータの利用が必須となる (*Journal of the Earth Simulator* 2006)。我々のグループは地球シミュレータをゲノム配列解析に利用することが認められている唯一のグループである。作成した大規模 BLSOM 上へ、大学や企業を含む研究機関等の各研究者が興味を持つ配列を、各自が PC レベルの計算機を用いてマップすることで、それらの生物系統や遺伝子機能が推定できる。その目的の大規模 BLSOM の公開を予定している。

＜達成できなかったこと、予想外の困難、その理由＞

地球シミュレータで認可された計算時間上での制限のために、5 連塩基頻度を用いた BLSOM については、一部の生物種に限定した予備的な解析しかできなかった。

＜今後の課題＞

体内環境由来の試料に関してのメタゲノムデータが急増する傾向にあるので、BLSOM マップデータの更新を続ける。BLSOM で得られた系統推定の大量な結果からの知識抽出を医学分野の研究者と共同で行なうことを計画している。

タンパク質の機能推定に関しては、微生物ゲノムに由来するタンパク質の機能推定を中心に解析を進めてきた。しかしながら、医薬学的には高等生物のタンパク質の機能推定も重要であるので、それらをも今後の解析に含めたい。

DNA シークエンサーの性能の飛躍的な進歩で、250nt レベルの配列なら大量配列が比較的安価で解読が可能になり、メタゲノム解析へも導入される傾向にある。250nt レベルの短い配列の系統推定を可能にする BLSOM 法の開発を行なう。

＜成果公表リスト＞

1) 論文／プロシーディング（査読付きのものに限る）

1. 0801251138

Abe, T., Sugawara, H., Kanaya, S. and Ikemura, T.: Characterization of genetic signal sequences with Batch-Learning Self-Organizing Map (BL-SOM). Proceedings of Workshop 2007 on Self-Organizing Maps, 2007.

2. 0701121230

Abe T., Sugawara S., Kanaya S., and Ikemura T., 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.

2)特許

特許番号 3928050、池村淑道、阿部貴志、上月登喜男、金谷重彦、木ノ内誠：発明の名称「塩基配列の分類システムおよびオリゴヌクレオチド出現頻度の解析システム」：特許確定日、平成 19 年 3 月 16 日