B1

自己組織化地図によるゲノム情報の包括的視覚化

総合研究大学院大学　葉山高等研究センター　○池村　淑道

Comprehensive visualization of genome information on a self-organizing map (SOM).

Toshimichi Ikemura, The Graduate University for Advanced Studies

Abstract:
With the increasing amount of available genome sequences, novel tools are needed for comprehensive analysis and visualization of species-specific sequence characteristics for wide varieties of genomes. An unsupervised neural-network algorithm, Kohonen's self-organizing map (SOM), is an effective tool for clustering and visualizing a large amount of complex data on a single map. We modified the conventional SOM for genome informatics, making the learning process and resulting map independent of the order of data input. We used the modified SOM to characterize di-to pentanucleotide frequencies in a total of approximately 10-Gb sequence derived from both prokaryotic and eukaryotic genomes for which complete sequences are known. SOMs could classify the 10- and 100-kb sequences of these genomes mainly according to species on a single map and revealed sequence characteristics of individual genomes. The unsupervised algorithm could recognize, in most of the sequence fragments, the species-specific characteristics (key combinations of oligonucleotide frequencies) that are signature features of each genome. In other words, SOMs could systematically extract profound genomic information from the oligonucleotide frequency in each genome. Because species-specific separation on a SOM was very clear, SOM could be established as a novel strategy for phylogenetic classification of sequence fragments obtained from uncultured microorganism mixtures in an environmental or clinical sample.

1. はじめに

ゲノム塩基配列の解読は益々加速する勢いにあり、既に200を超えるゲノムの完全配列が解読されている。広範囲な生物のゲノム配列の多様性を統合的に理解することは、ゲノム科学の重要な課題である。我々は、コホネンが記憶やその想起の機構を研究するために開発した自己組織化マップ(SOM)に着目した。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させている。応用性の広い方法として、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。出来上がった地図がデータの入力順や初期条件に依存することは、ゲノム配列の解析では不都合であった。『学習過程と結果の地図構造がデータの入力順序に依存しないようにする』という新しい特徴をSOMに導入し、初期条件に主成分分析の結果を反映することで、強力なゲノム情報解析と可視化のtoolとして確立できた。

2. 研究開発項目とその成果概要

2. 1 革新的なゲノム情報解析技術としての自己組織化地図(SOM)

SOMは教師なしニューラルネットワークアルゴリズムであり、大量情報の全体像と部分情報の両方を

効率的に把握できる。本研究開発では、従来型のSOMを、データの入力順に依存しない一括学習型SOMに変更することで、ゲノム情報解析の革新的な技術として確立した。配列解読の完了した原核と真核生物のゲノム類について、連続塩基の出現頻度を解析することで、ゲノム断片配列の大半を生物種により分類できた(Genome Res. 2003)。このゲノム解析技術の革新性を、環境由来のゲノム配列を対象にした解析例で説明する(図1)。環境中の微生物類は、実験室で培養が困難な例が大半であり、膨大なゲノム資源が未開拓に残されてきた。培養を行わずに、混合状態のままゲノムDNAを抽出し、DNA断片を配列決定する研究が開始されている。科学的・産業的に重要で有用な新規遺伝子類を発掘する手法として期待されている。例えば、Venterらはバーミューダ沖のSargasso海の微生物集団から混合ゲノムDNAを回収し、80万本の断片配列を決定し約120万の遺伝子の候補を推定した。このような環境由来の大量ゲノム配列を対象に、生物系統の推定や多様性の実体を知ることが重要となった。連続塩基の出現頻度のみで生物種に分類ができる教師なしアルゴリズムのSOMは、新規配列類の系統推定において、オルソログ配列セットや配列間のアラインメントが不必要であり、新規性の高い配列類の系統分類には最適な方法である。この目的を実現するために、図1ではDNAデータベースに収録されている約1500種の既知原核生物種由来の総計1.5Gbの配列を5kbに断片化し(1kbでも良いが分離能はやや下がる)、4連続塩基(tetranucleotide)の出現頻度についてSOMを行った。

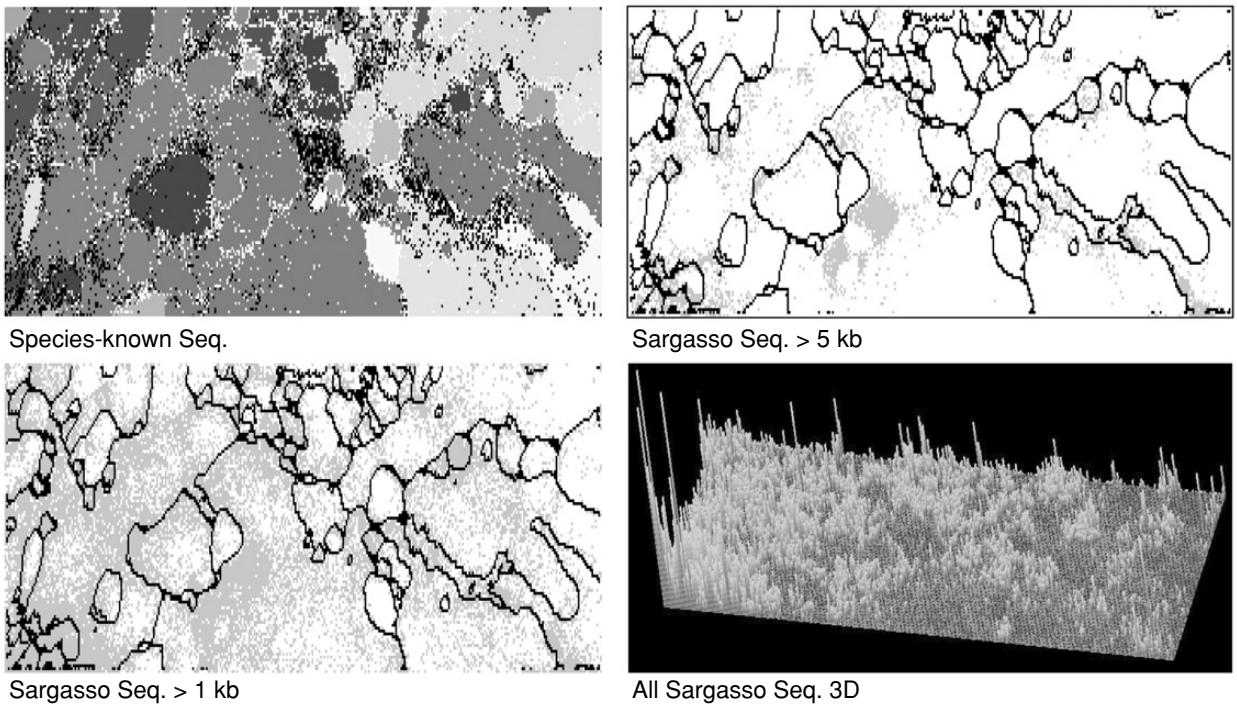

図1

難培養性の環境微生物類の配列を解析する場合、未知な生物種に由来する可能性が高く、どの系統に近いのかを推定することが重要になる。図1のSpecies-known Seq.では、約1500種の既知原核生物のゲノム配列について、25の系統群への分類の様子を表示しており、約85%の配列が正しい系統を反映して分離していた(解析法や表示方法は、Genome Res., 13, 693-702, 2003で説明している)。Venterらが報告している大量の断片配列を、このSOMへマップすることで、どの系統に近い配列がどのような量比で混在しているのかを推定できる。図1のSargasso Seq.>5 kb (5 kb以上のcontig配列をマップした)では、環境中の優先種の配列を解析でき、Sargasso Seq.>1 kbやAll Sargasso Seq.3Dでは、少量しか存在しない生物種由来の配列

が解析できる(3Dの意味は下記の図2で説明する)。新規性の高い大量な配列類を系統分類し、ゲノムごとにin silicoで再構成が可能となった。オロソログ配列セットや配列間のアラインメントが不要であり、系統推定の従来法からは予想できない革新的な技術である。

2. 2 cDNA配列を対象にした、ゲノム機能領域のSOM解析

ゲノム機能の解明、特に遺伝子の発現制御機構を研究する場合には、非翻訳領域の役割を知り、配列上の特徴を明らかにすることが重要となる。理研で配列決定がなされたマウス完全長cDNAを対象にした、SOM解析の結果を例に、機能解析におけるSOMの有用性と革新性を紹介する。次ページの図2Aでは、約4万本のマウスの完全長cDNAを対象に、4連続塩基頻度をSOM解析した。タンパク質をコードするcDNAのみからなる格子点を紫で、コードしないncRNAのみからなる格子点を赤で表示した。色付き棒の高さは、各格子点(ニューロン)に帰属した配列の数を示す(図1の3Dの意味)。連続塩基頻度以外の情報を与えていないのに、SOMはタンパク質をコードするcDNAとncRNA配列をほぼ完全に分離している。分離の原因として、タンパク質をコードするCDS領域からのコドン使用の影響が考えられるので、図2Bでは、タンパク質をコードするcDNAについては5'UTR・CDS・3'UTRの3領域に分割し、ncRNAを含めた4カテゴリーについてSOMを行い3D表示した。連続塩基頻度以外の情報を与えていないのに、機能領域による明瞭な分離が起きており、SOMが各機能領域の特徴を的確に識別することを示している。複数のカテゴリーの配列が混在する格子点では、各機能カテゴリーの配列数を対応色の棒で積み上げている。各配列に関するアノテーションデータとリンクしており、AVSを基礎にしたズーム機能を備えている。積み上げ棒の特定部分を指定することで配列やアノテーション情報を瞬時に知ることができ、能率的な知識発見を可能にする。

2. 3 遺伝子発現プロファイルのSOM解析

マイクロアレイ実験技術の進展に伴い、一つの生物についても数千-数万の遺伝子に対する発現プロファイルを同時に測定することが可能で、数百の異なった実験条件についてゲノム全体の遺伝子発現プロフ

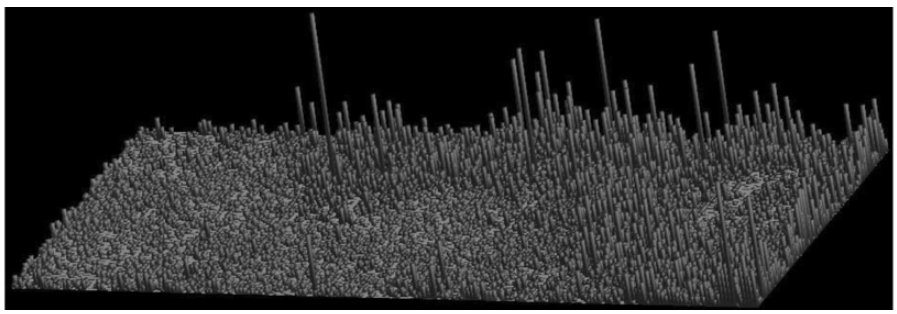

図2A Mous完全長cDNAについてのTetranucleotide-SOM
protein-coding cDNA : ncRNA

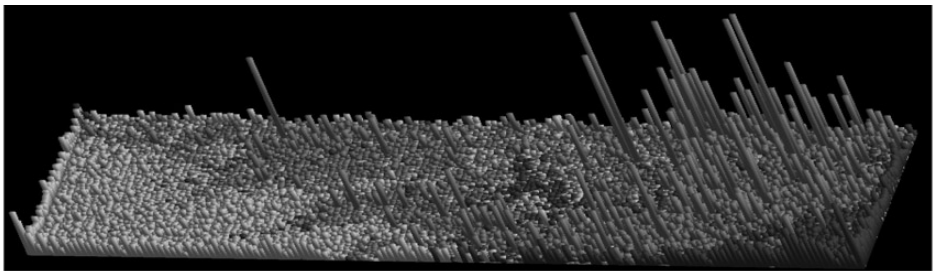

図2B 5'UTR : CDS : 3'UTR : ncRNA

ァイルが得られる。複数の実験条件で得られたプロファイルの類似性から遺伝子を分類することで、発現プロファイルが類似な遺伝子を探索するだけでなく、遺伝子発現の組織ならびに環境の特異性と関連した制御メカニズムをゲノム規模で理解することが可能となる。SOMを用いて、複数条件下でのマイクロアレイの結果を解析することで、遺伝子を高精度で分類できた。具体的には、マイクロアレイ実験により経時的に測定された発現プロファイルデータについて、細胞あるいは組織全体における遺伝子発現の変化を効果的に可視化し把握するための有用性の高い方法を確立できた。

3. ネットワークの活用について

SOMの解析結果は大規模計算から得られる貴重な情報資源であり、ネットワークを介した利用を可能にしている。既存の配列で作成した多種類のSOMから、利用者側が目的に合ったSOMを選択して、新規配列をマップし、系統や新規性の推定を行うことが可能である。各変量（各連続塩基）の成分マップからも重要な情報が得らえる。5連塩基の場合には、1024枚のマップを参照することになるので、それらを管理し利用を容易にするシステムが必要であり、塩基配列データについてのアノテーションデータとのリンクも重要となる。ネットワークを介した公開を目標としているので、公開用に便利なPostgresでシステムを管理運用している。

4. まとめ

SOMを、当初の予想よりは遥かに強力なゲノム情報解析と可視化toolとして確立できた。

5．研究開発実施体制

代表研究者　総合研究大学院大学　葉山高等研究センター　池村　淑道

研究分担

研究開発項目：真核生物ゲノムのSOM解析

総合研究大学院大学　国立遺伝学研究所　池村　淑道、深川竜郎

研究開発項目：SOMによるゲノム情報の統合と視覚化の基本アルゴリズムの開発

奈良先端科学技術大学院大学　情報科学研究科　情報生命学専攻　金谷重彦

研究開発項目：連文字解析用SOMプログラムの構築と評価

山形大学　工学部　応用生命システム工学科　工藤喜弘、木ノ内誠