公募研究：2003～2004年度

ゲノムに潜むシグナル・モチーフ部品の網羅的探索のための自己組織化地図

●池村　淑道

1）国立遺伝学研究所進化遺伝研究部門 (2003年度)、総合研究大学院大学　葉山高等研究センター（2004 年度）

〈研究の目的と進め方〉

大量なゲノム情報から未知の基本的な知識を得ることは、生命情報科学の重要な課題である。コホネンが記憶やその想起・連想の機構を研究するために開発した自己組織化マップ法(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)ことを実現している。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。以前から、奈良先端大・山形大のグループとの共同研究として、コホネンの従来型のSOMを、データの入力順に依存しない一括学習型SOMに変更し、ゲノム配列情報の解析技術として確立してきた。一括学習型SOMに変更したところ、予想を遥かに超える有用性を見出した。複数の生物種の断片的な(例えば10kb)ゲノム配列だけが与えられたのでは、どの生物の配列なのかを識別することは不可能に思える。しかし、各々のゲノムにはオリゴヌクレオチド(例えば4連続塩基)頻度に関する個性が内在しており、そのゲノムの個性を識別して、付加情報がなくても断片配列を生物種ごとに分類（自己組織化）可能であった。

ゲノム上には、シグナルやモチーフ配列と呼ばれる遺伝子発現において重要な役割を担う、情報上の部品類が多様な組み合せで存在する。本研究では広範囲の生物種を対象にシグナルやモチーフ類を情報学的に抽出する手法を確立する目的で、SOMの持つ高いクラスタ分離能を基礎に、4-8連塩基の頻度をSOM解析する。実験的な研究の進んでいる生物種を対象に、既知のシグナルやモチーフ配列の出現頻度パターンについてSOMによる特徴抽出を行っておき、配列は解読されたが他の実験的な研究の進んでいないゲノムに関して、*in silico* のシグナル配列の探索法を開発し、並行してGenomeWordDictionaryと呼ぶ情報部品のデータベースを構築する。

多様な環境で生育する微生物類は培養することが困難な例が大半を占めており、通常の実験的アプローチが困難であったため、膨大なゲノム資源が未開拓に残されてきた。新規性の高い遺伝子類を豊富に保有すると考えられ、科学的のみならず産業的にも注目を集めている。これら難培養性微生物類を解析する新技術として、環境中の生物集団の混合試料から培養操作なしにゲノムDNA混合物を抽出し、断片化ゲノム配列をクローン化し、配列決定を行い、遺伝子探索を行なう技術が開発され、世界的に普及をはじめている。多様な環境に生育する生物種の全体像の把握を可能にする有力な方法である。しかしながら、新規性の高い遺伝子配列ほど、配列相同性検索が適用できず、どの生物系統に属し、どれだけ新規性の高いゲノムに由来していたのかを推定することが困難である。ゲノム配列解析用に改良を進めてきたSOMは、連続塩基の出現頻度の類似度のみで、断片ゲノム配列を生物種ごとに分離（自己組織化）する能力を持ち、この目的に最適な方法である。既知の微生物種のゲノム由来の全断片配列を対象に、高性能スーパーコンピュータを用いて大規模SOMを作成し、更新を続け公開して行けば、各研究者が配列解読を行った環境由来の遺伝子配列を、PCレベルの計算機でこの大規模SOM上にマップすることで、配列の由来する生物系統や新規性を各自が推定できる。

〈研究開始時の研究計画〉

2003年度の研究の当初計画

1）塩基基配列が解読されたゲノム全体を対象にした、4-8連塩基頻度のSOM解析。ゲノム配列が解読されたゲノムを中心に、100種類以上のゲノム配列の全体に関して、4-8連塩基頻度のSOM解析を行い、各ゲノムにおいて特徴的な頻度で出現する連続塩基配列を網羅的に探索して、機能上の意味との関係を知る。

2）GenomeWordDictionaryの構築。SOM解析で得られた特徴的な連続塩基配列の生物学的な意味を知るためには、各配列について、実験的な研究を報告した文献類を組織的に参照することが重要になる。この文献検索の過程で蓄積する検索情報自体も有意義なデータセットとなる。SOMでの知識発見と統合して、連塩基配列に関するDictionaryを作成する。

2004年度の研究の当初計画

1）SOMによるゲノムの機能領域の特徴抽出。ゲノム配列のみならず、cDNA配列をも対象にして連続塩基頻度のSOM解析を行い、各ゲノム上の機能領域の特徴を明らかにする。

2）SOMの難培養性微生物類の混合ゲノム解析への適用。SOM法を用いて配列相同性検索に依存しない、ゲノム断片配列の系統推定法を確立する。オロソログ配列セットの存在しない、新規性の高い遺伝子配列の系統推定が可能になる。

3）GenomeWordDictionaryの公開用の整備とSOMデータの高機能可視化。

〈研究期間の成果〉

2003年度の研究成果

1）塩基基配列が解読されたゲノム全体を対象にした、4-8連塩基頻度のSOM解析。ゲノム配列の解読が進んでいる約150種類のゲノム配列の全体に関して、4-6連塩基頻度のSOM解析を行い、各ゲノムにおいて特徴的な頻度で出現する連文字配列を明らかにした(1,2)。特に、回文型の配列類は生物種の特徴を顕著に反映する傾向にあった。8連続塩基については回文型、7連続塩基については中抜きの回文型の配列についてのSOM解析を行ったところ、次元数を低くおさえていながらも、多様な特徴抽出が可能になった。シグナルやモチーフ配列の多くが回文型であることに関係すると考えられる。1kb程度のヒトやマウスの断片配列をSOM解析すると、単一のゲノムについても、5’と3’UTR、CDS、イントロン領域等で明瞭に分離する傾向を示した。さらに、上記の各機能領域内部についても分離しており、機能の細分化と関係するシグナル配列を探索する新規な情報学的手段を提供すると考

えられる。シグナルやモチーフ候補群が特定の組み合せで集中するゲノム部位の探索が行え、その組み合せの機能上の意味を検討することが可能になった。
2)GenomeWordDictionaryの構築。各連続塩基についてSOMの画像データを収録し、着目する連続塩基に関する実験的研究を報告している論文名とAbstract等を収集し、『ゲノム語辞書；GenomeWordDictionary』と呼ぶ新規な辞書を編纂し、世界へ発信するためのシステムを構築した。論文の文献データについては、各連続塩基配列に関するPUBMEDの検索結果について、MEDLINE形式で収録している。現時点では、4連続塩基の全体について収録を完了した。ATGCの４文字からなる通常の言語辞書形式であり、誰でも容易に参照できる。関係データベースであるので、生物種や系統ごとにも辞書が作成できる(例えば、HumanGenomeWordDictionary)。通常の辞書形式で各連続塩基配列別に文献類が収録されているので、転写因子への結合配列を代表例とする遺伝シグナルに対応する連続塩基配列類の機能的な意味を効率的に把握でき、生物学的な意味の特定が可能になっている。広範囲の生物種の多様なシグナルやモチーフ配列を、辞書形式で集大成できれば、シグナルやモチーフ配列を代表例とする機能配列に関して実験家の得ている知識の全体を容易に把握でき、配列決定以外の分子生物学的な研究が進んでいないゲノムについて、シグナルやモチーフ群のin silico探索が可能になる。SOMが明らかにした、各ゲノムを特徴付ける連続塩基配列類の生物学的な意味を知る上でも必須のデータベースである。

2004年度の研究成果

1) SOMによるゲノムの機能領域の特徴抽出。マウスの約４万本の完全長cDNAについて、5-6連塩基の頻度をSOM解析したところ、protein-codingとprotein-noncoding cDNAで分離する傾向にあった。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効果が考えられるので、protein-coding cDNAについては5’UTR・CDS・3’UTRの3領域に分割し、protein-noncoding cDNAを含めた4カテゴリーについてSOMを行ったところ、連続塩基頻度以外の付加情報を与えていないのに、4カテゴリーによる明瞭な分離が起きており、各機能領域を特徴付けるシグナル配列類を抽出することが可能になった。
2) SOMの難培養性微生物類の混合ゲノム解析への適用。環境中で生息する微生物類の大半は実験室で培養が困難であり、未開拓なゲノム資源として残されてきた。培養せずに混合ゲノムDNA試料のショットガンシークエンシングを行う方法が普及してきた。教師なしアルゴリズムのSOMは、生物種に関する予備知識なしに断片配列の大半を生物系統に分類可能であり、オロソログ配列セットの存在しない新規性の高い遺伝子配列の系統推定が可能になる。この目的を実現するために、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連続塩基の出現頻度についてSOMを行った。上述の既知原核生物のゲノム配列に関して、25の系統群への分類を解析したところ、約85%の配列が正しい系統を反映して分離していた。Venterらが報告している大量の断片配列を、そのSOM上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たのかを推定できた（3)。
3) GenomeWordDictionaryの公開用の整備とSOMデータの高機能可視化。GenomeWordDictionaryを、当初は関係データベースシステムとしてオラクルを用いていて構築してきたが、公開用として利用者側に制限のない使用を可能にする目的で、ポストグレスへの移植もおこなった。SOMの結果についての多数の画像データを、AVS機能を用いてVR技術を基礎にした立体視を含む高機能映像化を行い、配列に関するアノテーションデータと連結させた。

〈国内外での成果の位置づけ〉

一括学習型のSOMを世界に先駆けてゲノム塩基配列の解析に導入し、ゲノムインフォマティクスの革新的な技術として確立してきた(1,2)。難培養性微生物類の混合ゲノム解析は、米国を中心に国策的に大規模な塩基配列の解読が行われている。この混合ゲノム解析へ、SOMは強力で革新的な情報解析技術を提供できる(3)。地球シミュレータを使用し、現時点で公的データベース登録されているほぼ全てのゲノム配列を一枚の５連続塩基SOM上にマップすることを可能にした。世界に類例のない先進的な解析方法であり、多様な環境に由来する微生物ゲノムを研究している複数の実験グループからの依頼で共同研究が進行している。日本工業新聞（2003／4／７）に紹介記事が掲載された。日本学術会議　遺伝学研究連絡委員会　合同シンポジュウム『これからの遺伝学』（2003／8／7）で、発表をおこなった。

〈達成できなかったこと、予想外の困難、その理由〉

次元数と計算時間の制限のため、7と8連続塩基につては、回文型を中心としたSOMしか行えていない。GenomeWordDictionaryはポストグレスシステムを用いており、ネットワークを介して自由に利用できるシステムになっているが、セキュリティ等に関する技術に不安があり、一般への公開が行なえていない。

〈今後の課題〉

SOMを用いたタンパク質機能推定法の開発。広範囲のゲノム配列が解読された結果、アミノ酸配列の相同性検索では機能が推定できない、機能未知なタンパク質が大量に蓄積し、産業的にも未利用なまま残されてきた。アミノ酸の１次元配列の相同性検索に依存しないタンパク質の機能推定法が求められている。タンパク質の機能には､アミノ酸の１次元配列よりは3次元構造が重要なので、１次元配列の相同性検索ではなく、構造や機能モチーフを含む連続アミノ酸の使用頻度に着目した機能推定法を開発することが可能である。2～5連続アミノ酸頻度のSOMに着目している。

〈研究期間の全成果公表リスト〉

１）論文／プロシーディング（査読付きのものに限る）

1. 0308291550 Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. Informatics for unveiling hidden genome signatures. Genome Res., 13: 693-702 (2003).
2. 602011429 Abe, T., Kanaya, S., Kinouchi, M., and Ikemura, T. Genome informatics for unveiling hidden genome signatures Proceedings of the Institute of Statistical Mathematics 52: 207-215 (2004)
3. 602011443 Abe, T., Ikemura, T., Kanaya, S., Kinouchi, M., and Sugawara, H. A novel bioinformatics strategy for phylogenetic study of genomic sequence fragments: Self-Organizing Map (SOM) of oligonucleotide frequencies Proceedings of Workshop 2005 on Self-Organizing Maps WSOM2005, 669-676

(2005)

2) データベース／ソフトウェア

『ゲノム語辞書；GenomeWordDictionary』

3) 特許など

国際特許出願PCT/JP2004/002771「塩基配列の分離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」