公募研究：2000～2001年度

# 二次元色彩化によるゲノムデータベースの解析

●大澤　研二[1)]　◆尾畑　伸明[2)]　◆吉田　徹彦[3)]　　◆大沢　健夫[1)]

1) 名古屋大学大学院多元数理科学研究科　　2) 東北大学大学院情報科学研究科　3) （株）東亞合成名古屋総合研究所

**〈研究の目的と進め方〉**

私たちは一次元配列中に存在する反復や特定の記号の分布の偏りなどを見い出す解析手法として二次元色彩化を開発した。この方法では、既知のデータとの類似性に依存するのではなく、配列自身に含まれる規則性を引き出すことが可能となり、大腸菌全ゲノム中のタンデムリピートの発見や数字列などに潜在する規則性の顕在化において、その実用性が証明されている。

本研究では、二次元色彩化を用いて様々な生物のゲノムデータから新しい知識を発見する方法を確立することを目的とする。そのために、色彩化の条件設定の検討を行い、さらに、一次元の配列から二次元の色彩パターンを作り出すための処理の自動化を試みる。これによって、様々な生物のゲノム情報から得られたデータを比較することが可能となり、未知の傾向を有した遺伝子やタンパク質の発見を促進できるものと期待される。

一方、パターンの解析には視覚による判別を採用しており、その基礎となる原理に関しても明確にはできていない。この問題点を解消し、さらに多くの有用な情報を得るためには、パターンの数値化が必要となると考えられる。ここでは、手法の自動化を視野に入れて、現在の解析方法に基づいた実際のデータ解析を進めるとともに、パターン自身の解析についても研究を進める。

**〈研究開始時の研究計画〉**

本研究では、二次元色彩化の応用範囲を広げるとともに、手法改良に重点を置いて研究を進める。そのために、大きな研究項目を4項目設定する。

1. ゲノム情報中の新しい特徴を持つ配列の発見のために、二次元配列化における幅の設定を、連続的・不連続的に変化させたものを並べおくだけでなく、固定幅に関しても具体的な例を用いて検討する。
2. 配列中の特徴を効率良く顕在化させるために、色彩化における塩基あるいはアミノ酸残基に応じた色の設定の問題の検討を行う。
3. データの比較のための色彩化パターンの分類法の確立を目指して、視覚によって認識される分類要素の抽出を試みる。
4. 様々な操作を効率的に行うためには、二次元配列化の自動化が重要となるので、反復配列探索を目的として開発されたプログラムの改良を行う。

以上の計画を遂行するとととともに、従来の方法を用いた各種生物ゲノム中のタンデムリピートの探索についても継続して研究を進め、それらの配列や位置関係を比較することによって生物種間の進化的関係の分析が可能となるかという点についても解析を進める。

**〈研究期間の成果〉**

２０００年度は二次元色彩化を用いて、ゲノム中のタンデムリピートを探索する目的で開発した方法を、インフルエンザ菌に適用し、30塩基以上の反復長のタンデムリピートを30個見出した。インフルエンザ菌では大腸菌とは異なり、100塩基前後の長さのリピートは割合として多くないなどの特徴を示した（成果リスト１）。一方、ゲノム全体に関しての特徴を見出すために、表示幅を一定にして単一の彩色図を作成する手法を用いて、様々なゲノムを解析した。すると、細菌において、それぞれに特徴的な色パターンを示しているのに加えて、枯草菌ゲノムではGC相対比などに応じたパターンが顕在化したのに対して、大腸菌ゲノムではそれが明確にならなかった。これは、色彩化において、それぞれの塩基の種類に対応する色彩の選択が特徴の顕在化に重要な因子となることを示している。

２００１年度は下記の５テーマについて成果を挙げた。

1．ヒト21番染色体中のタンデムリピートの解析：ヒト21番染色体中の28Mセグメント中に30塩基長以上の周期をもつタンデムリピートを約300ヶ所同定した。もっとも長い周期は639塩基であった。ある領域に集中する傾向やかなりの長さに渡ってタンデムリピートが存在しない領域が見られた。

2．多くの細菌ゲノム中のタンデムリピートの探索：弘前大学の清水俊夫教授との共同研究により、大腸菌以外の細菌についても探索を行い、多くの知見を得た。特に一部の細菌ゲノムでは、非常に多くのタンデムリピートの存在が明らかとなった。また、ピロリ菌のゲノムでは、大きなinverted repeatの存在が明らかとなった（成果リスト２）。

3．他の探索手法との比較：Bensonの開発したTandemRepeatFinderとの比較により、二次元色彩化でのみ発見できるタンデムリピートの存在を確認した。ヒト21番染色体のRepeatMaskerによる反復配列の解析結果と比較し、多くのタンデムリピート（二次元色彩化により発見されたものの約40%）が二次元色彩化でのみ見出されることを確認した（成果リスト２）。

4．長周期タンデムリピートの探索：特に長い周期を持つタンデムリピートに関しては、ただ一つの幅のパターンを検証すればよいと考えられることを確認した。

5．視覚による判断の基準の個人差検定：ゲノム配列中のタンデムリピートの存在に起因するパターンの発見において、それに携わる人間の個人差による違いの有無を調べるために、同じ細菌ゲノムの配列を対象としたタンデムリピートの探索を三人の研究者によって実施し、その差を検定した。ある程度の差が生じることは認められたが、全体として問題となるほどのものではないことが確認された（成果リスト２）

また、この二次元色彩化法に関して、特許を申請し、日本、アメリカ、ヨーロッパのそれぞれの特許として認められた（成果リスト３，４，５，６）。

**〈国内外での成果の位置づけ〉**

ゲノム情報を色彩化により解析する方法の開発について、国外で類似の研究を行っているところはない。国内では、産業技術総合研究所の鈴木理研究員のグループが行っており、特に塩基種類の分布の偏りに注目しているようである。タンデムリピートの探索は異なる手法を用いるグループが幾つかあり、それぞれに成果を出している。しかし、手法の開発を主に行い、様々な生物種のゲノムの解析を行った例はほとんどない。

また、この方法を基にした大規模計算実験が東亜合成とNTTデータの協力によって実施され、かなりの成果を挙げている。これらの手法は特許として認められたことから、今後さらに産業的な応用に向けて、ゲノム情報解析に留まらず、様々な情報の解析への応用など、注目を集めるものと期待される。

**〈達成できなかったこと、予想外の困難、その理由〉**

処理の自動化はかなり進み、使い勝手の良いものが出回るようになったが、その一方で、最終段階までを自動化する手法の開発は様々な問題から、達成できなかった。特に、個人差から来る問題の解決に向けて、パターン抽出を目指したパターンの数値化を試みたが、その展開はうまく運ばなかった。結果的には、画像として提示されるパターンの特徴を分類する段階で、人間の視覚が行っている判断の基準を引きだすことができなかったことが最大の困難であった。このような経験に基づき、パターン認識との兼ね合いから、さらなる特徴抽出のための検討が必要と考えられた。

一方、タンデムリピートの探索については、かなりの数の細菌種のゲノムの解析が弘前大学の清水教授との共同研究により進んだが、その結果をまとめるには至らなかった。当初の計画にもあった生物間の進化関係への応用についても、これらのデータを基に、様々な比較検討が行われたが、有用な知見を導きだすには至らなかった。ゲノムの編成過程におけるタンデムリピートの役割が様々な研究で報告されているが、そこに統一的な規則が存在するのではないかとする仮定は、今のところ裏付けがとれていない。原因としては、タンデムリピートの中にも様々な役割をもつものが存在し、それらを細かく分類する必要があるためではないかと考えられる。今後も、これらの比較検討を継続し、生体内での機能とゲノム編成における機能の両面に渡っての解析を進める必要があると考えられた。

**〈今後の課題〉**

一次元の情報配列の解析手法として、二次元色彩化の有用性は十分に証明されたので、今後はこれらの解析処理を自動化することが急務と考えられる。特に、二次元のパターンとして提示されたものを計算処理によって抽出する作業をどのように行うかについて、基準の設定など、さらに基礎的な研究を進めていくことが必要となるだろう。

一方、ゲノム情報解析においては、従来のように既知情報が少ないことから生じる問題はかなり減ってきたとはいえ、まだこれらの問題が様々な場面で出てくるものと考えられる。さらには、従来のようにタンパク質に発現される情報のみを処理していたのでは、生命活動を解明することは困難であろうから、それらの情報についても、二次元色彩化によるパターン抽出が何らかの形で貢献するところがあるように思われる。

さらに、情報社会において、ゲノム情報のみが対象となるわけでもないから、記号列を用いるものであれば、何にでも応用可能であると思われる。そのような考えに基づき、たとえば、文学作品や手紙などの作者や真贋の鑑定などについても、応用できる可能性はあるものと考えられる。これらの点をふまえて、その他の分野にも対象を広げて研究の幅を広げていく必要があると思われる。

**〈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．Yoshida, T., Obata, N., Oosawa, K.: Color-coding reveals tandem repeats in the Escherichia coli genome. J. Mol. Biol., 298, 343-349 (2000).

2．Mizuta, S., H. Munakata, A. Aimaiti, K. Oosawa, T. Shimizu Evaluation of the color-coding method for searching tandem repeats in prokaryotic genomes. Chem-Bio Informatics J., 4, 133-141 (2004).


**3）特許など**


3．吉田徹彦、尾畑伸明、大澤研二、記号列の特徴顕在化方法、日本特許番号第３１４９８２４号　(2001)

4．T. Yoshida, K. Oosawa, & N. Obata,　Method and apparatus for revealing latent characteristics existing in symbolic sequences, United States Patent, US6438496 B1 (2002)

5．T. Yoshida, K. Oosawa, & N. Obata,　Apparatus for manifesting latent characteristics existing in sequences of symbols, United States Patent, US2002172971 (2002)

6．T. Yoshida, K. Oosawa, & N. Obata,　Method and apparatus for manifesting characteristic existing in symbolic sequence, European Patent, EP1501025 (2005)