プロジェクト名：高速相同性発見手法を用いたゲノム解析とその応用

１．プロジェクト体制

研究代表者
［国立情報学研究所］　宇野　毅明
共同提案者
［国立遺伝学研究所］　小出　剛
（研究協力者）
［国立遺伝学研究所］　梅森　十三

２．これまでの研究成果及び今年度の提案内容

2-1. これまでの研究成果

マウス 13 番染色体上の *Genic1* 領域は、複雑な類似性（繰り返し配列）を多く持つため難解読領域となっている。そこで、正確な配列解読を目的として、既読の BAC 配列の再アセンブリを行った。この結果、今まで不明瞭だったギャップが生じる原因と、実際のギャップ領域の位置の推定に成功した。これらの領域の多くは、数百塩基の繰り返し配列の近傍に存在しており、シークエンスアセンブル時に生じる BAC 配列間の矛盾によりギャップが生じていると考えられた。また、これらの小さなギャップを小規模なエラーと見なし、BAC 配列のアセンブルを続けた結果、どうしても埋めることができないギャップを１つに絞り込んだ（図 1A）。このギャップは、ギャップ近傍の BAC 配列が間違っているか、２つの Contig 間を埋める BAC の配列が作られていない、或いはその配列が未だ読まれていないことから生じていると考えられた。平成 20 年度は、前者の可能性を考え、読み取りが不完全と思われる部分に対して実際に実験を行って配列の検証を行った。しかし、AC192334 という約 200kbp の BAC がデータベース上で見つかったことにより、このギャップの問題は解決した（図 1B）。つまり、宇野が開発した「高速相同性検索アルゴリズム」を用いて再びアセンブリを行った結果、AC192334 は Contig1 と Contig2 の橋渡しとなり、最後のギャップを埋めた。これらの

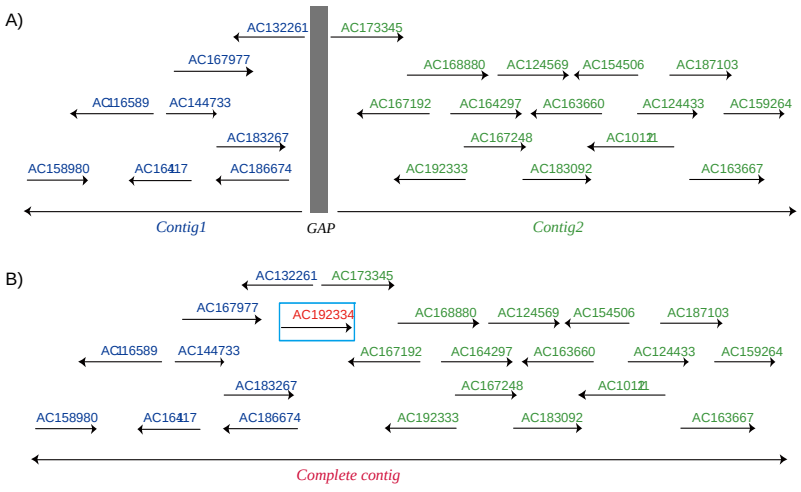
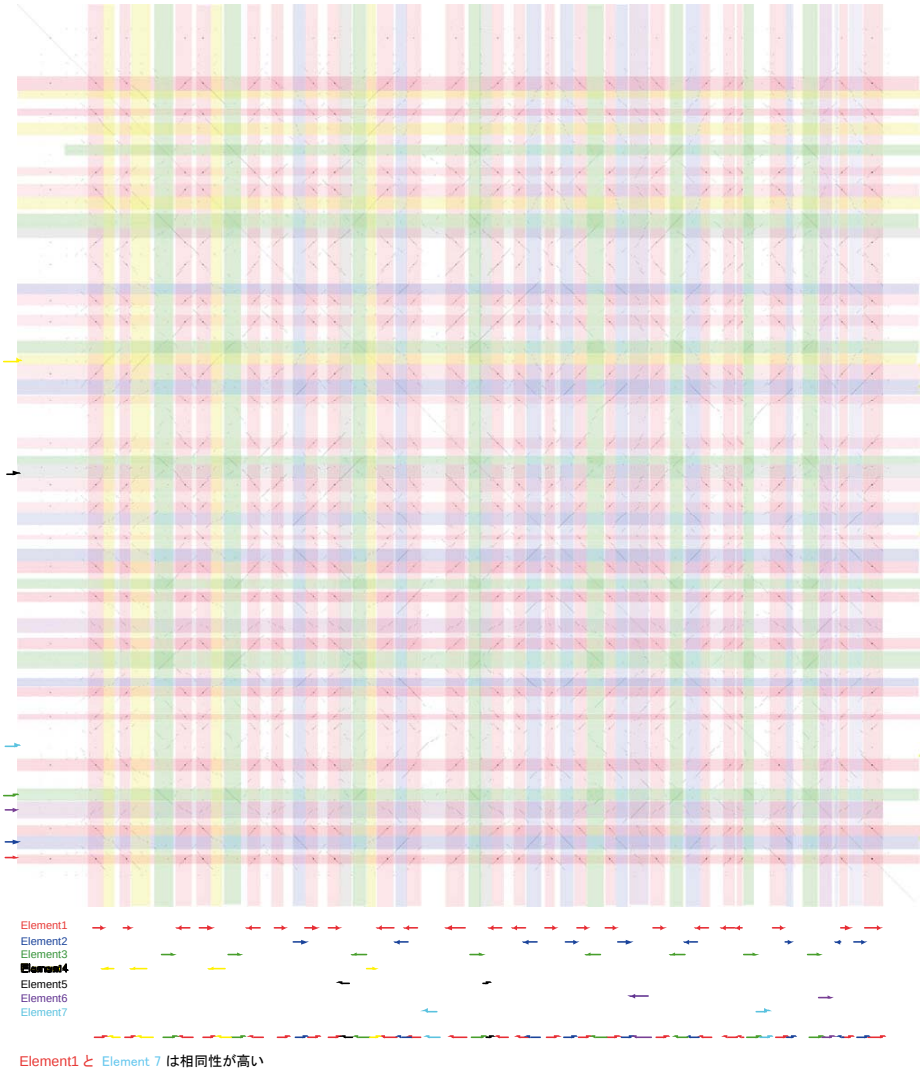



図 1 第 13 番染色体難解析領域の BAC コンティグ作成
A) 以前の BAC コンティグ
再アセンブルの結果、Contig1 と Contig2 にギャップがあることがわかった
B) 今年度に完了した BAC コンティグ
AC192334 をアセンブリに加えることで GAP が埋まり、完全に一続きになった contig が完成した

結果は、正しいBAC配列さえあれば、「高速相同性検索アルゴリズム」を応用するにより、正確かつ高速にアセンブリを行うことができることを示した。




（図2）第１３番染色体 **Genic1** 領域の繰り返し構造

平成20年度は、「高速相同性検索アルゴリズム」のさらなる応用として、複雑な類似性（繰り返し）構造を、視覚的に分かりやすく表現する方法を模索した。これは、繰り返し領域のエレメントを正確に知ると同時に、それが何回繰り返されているか等の全体構造を、視覚的により見やすくするのが目的である。このモデルとして、上述した正確にアセンブリされたマウス13番染色体の BAC contig 内で、特に繰り返し配列が多く存在する *Genic1* 領域(1.6Mbp)に対して解析を行った。「相同性高速検索アルゴリズム」により、２次元的に描かれた繰り返し構造の図を作成し、この情報から手動でエレメントの抽出を行った（図2、方法については中間報告書を参照)。これにより *Genic1* が持つ繰り返し構造を１次的に示すことに成功した。

2-2. 今年度の提案内容

1）高速相同性検索を用いたＤＮＡ配列のアセンブルシステムの確立とアライメントソフトの開発

本課題では、昨年度に引き続き「相同領域発見手法」により検出された相同配列の結果から直接アセンブリを行う方法の確立、さらには自動化プログラムの作成を目指す。これにより、作業を効率化できるだけでなく、将来的には高速でかつ繰り返し領域に強い、新たなアセンブリプログラムの開発が期待できる。4-1に示した通り、「高速相同領域発見手法」で描いた2次元の図を元に、繰り返し配列構造を手動で検出し、解析することができた。現在、正確な繰り返しエレメントの自動抽出を目指しているが、非常に困難を極めている。これは、目で見た場合には繰り返し配列の範囲をおおよそで掴むことができるが、自動化を試みた場合に、どの領域までが繰り返しなのかを判断するのが難しいためである。従って、この作業を完全に自動化することは難しく、マニュアルと組み合わせながら、正確な繰り返し配列の検出を進め、より効率的な手法の確立を目指す。

2）難解読領域の機能の探索

マウスやヒトのゲノム上には、*Genic1*のような複雑な反復領域が多数存在し、それらが完全なゲノム解読の支障となっている。そこで、ゲノム情報から、難解読領域を検索し、その領域について、本自動化プログラムが利用できるかどうか検討を加える。この解析により、本高速相同性検索アルゴリズムの有用性が確認できる。このような複雑な構造を取る領域は、セントロメアやヘテロクロマチンのように、そのゲノム構造が生物学的に重要な機能を持っていると考えられるため、解析が可能となる意義は大きい。また、*Genic1*を含めたこれらの特殊な領域を、ラットやヒトと比較することにより、ゲノム構造の進化や機能について、新たな知見を生む可能性も期待できる。

3．新領域融合プロジェクトへの発展の可能性

これまでの研究により、「高速相同性検索アルゴリズム」を応用することにより、従来ならば繰り返しが多くてアセンブリが困難であった BAC 配列について、高速かつ効率的にアセンブリを行うことができることが示された。現在は、この長い BAC 配列に加え、ショットガンシークエンス等の短い DNA 配列に対するアセンブリソフトのソフトの開発を目指している。しかし、「高速相同性検索アルゴリズム」を用いたアセンブルは、BAC の選定とアセンブリには強いが、大規模な相同性を小規模な相同性から決定するため、２つの配列のアライメントを細かく同定するような作業には向いていない。これには、小規模な相同性から中規模な相同性を決定づけるモデル、中規模な相同性から大規模な設定を自動的に決定するモデルの開発が必要である。この部分は、学術的に取り組むべき大きな課題である。

*Genic1* のような複雑な領域は、セントロメアやテロメア等の染色体「内」構造の維持だけ

でなく、核内の染色体配置等の染色体「間」構造の維持等、他にも生物学的に未知の機能を持っている可能性がある。従って、これらを含めたゲノム全体の正確な配列を得ることは、ゲノム科学の新たな知見を生み出す可能性がある。また、今回の成果の発展によって、マウスやヒト等の既読のゲノム配列に加え、今後加速的に解析が進むと考えられる他の生物種についてのゲノム配列の検証と、それによる正確な配列の決定が期待できる。これらを情報システム機構において公開するということは、ひとつの大きなプロジェクトの可能性であろう。
これらの技術を高めていくことで、相同性の計算を中心とした高性能ゲノム解析ソフトを開発するプロジェクトへとつながっていくことを期待している。

## ４．期待される効果等

上述の通り、近年繰り返し配列の重要性が認識されつつある。従って、繰り返し領域の生物学的な意味の解明は分子生物学や医学等多くの分野において重要な意味を持つだろう。また、繰り返し配列に強く高速なアセンブリソフトの開発は、ゲノム研究に携わる生物学者、情報学者全てにとって有用である。特に、萌芽的な研究を行うものにとっては、計算機コストの削減により、より多くの資源を研究に割くことができるようになる。アセンブリソフトの開発を基に、宇野が開発した「高速相同配列検索」が他の生物学的な解析、特に異なる種のゲノムの大域的な比較、繰り返し構造の解析、アセンブリのエラー検出などに応用されることを期待している。

## ５．予算金額及びその内訳

| 費　目 | 金額 | 主な使途 |
|---|---|---|
| 人件費 | 5,300 | 特別研究員の雇用 |
| 物件費 | | |
| 備品費 | 1,000 | 解析用コンピューター、データ用ハードディスク |
| 消耗品費 | 3,000 | シークエンス用試薬、BAC 購入費、プラスチック器具 |
| 旅　費 | 700 | 情報研、遺伝研双方への研究打ち合わせ旅費、学会への旅費及び参加費 |
| 謝　金 | | |
| その他 | | |
| | | 計　　10,000　千円 |

## ６．本課題の他の経費への応募状況

*Genic1* 領域の機能解析については、科研費・基盤研究 B に「マウスの遺伝子間相互作用異

常がもたらす不適合性の分子基盤とその進化的意義の解析」という課題名で申請中である。

７．その他