高速相同性発見手法を用いたゲノム解析とその応用

研究代表者：宇野 毅明

1. 共同研究者

［国立遺伝学研究所］　小出 剛、梅森 十三

2. 研究目標

2007 年現在、マウスゲノムは約 90%の塩基配列が決定されているが、未だに配列決定の為されていない領域が、全ゲノム中に広く点在している。これらの多くは、塩基配列そのものは決定されているが、データ解析が困難な領域である。一般的にゲノムの塩基配列は、DNA の断片化後に解読され、それらの配列はアセンブリによりその位置情報も決定される。しかし、そのようなデータ解析が難しい領域は、多くの場合ゲノム上に同じ並びが繰り返されている箇所（繰り返し配列）が存在するため、アセンブリが困難である。つまり、コンティグ作成時に、BAC 間のアセンブリで多くの矛盾が生じてしまう。これらの矛盾は、データベース上において、位置と長さが不確定であるギャップ領域として表現されている。

本テーマでは、すでに開発済みの高速相同領域発見アルゴリズムをこの領域に適用し、短時間で正確な配列の決定を行うことを目指す。また、アルゴリズム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、高速でかつ繰り返し領域に強い、Phred/Phrap/Consedに代わるような、新たなアセンブリプログラムの開発が期待できる。それとともに、マウスの未解読領域をシークエンスし、実際に解析を行いつつ並行してプログラムの開発を進めていくことで、繰り返し領域に関する困難性とそれに対するプログラムの性能を迅速に検証しつつ、同時に未解読領域の解明を行う。

この結果、これまでに解析が困難であった領域、つまり特定塩基の偏りや、繰り返し配列が高頻度に存在する領域を解析することができるようになると考えられる。これには、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ているセントロメアや、ヘテロクロマチン、高度重複配列領域などが含まれるため、解析が可能となる意義は大きい。このような配列をアセンブルにより得た上で、再び高速相同性検索アルゴリズムを用いて、反復配列や逆位配列などのゲノム上の特殊な構造について解析する。これにより、ゲノム構造の進化や機能について、これまでには無い新たな考察が可能になることも期待できる。

3. 平成19年度までの研究進捗及び主要成果物

〔研究進捗〕

平成 19 年度から開始。項６の「平成 19 年度の研究成果」を参照

4. 平成20年度以降の展開

・他のゲノム配列の決定

今回の研究で行った解析により、マウス 13 番染色体のギャップ領域を追加実験なしに補完することができた。この解析をマウスの 13 番以外の染色体のセントロメア、テロメア領域に代表される相同性の高い配列が繰り返し構造を持つ部分に対して行うことで、新たなギャップ領域の補完や、アセンブリの質の検証が行なえるだろう。これにより、追加実験なしに、既知のゲノムの質を高めていくことができるだろう。

・アセンブリングのモデル化

今回の解析手法をきちんと系統立てることにより、質の高いアセンブリング手法、および配列決定法を確立していくことができると考えている。

・実験も含めた総合的な、軽量なアセンブリング技術

現在、配列決定には多大な量のシークエンシングと、非常に重いアセンブリング計算が必要となっている。これを計算の高速化と高精度化により、実験の量と時間と手間を減らすことができると考えている。しかし、配列決定を低いコストで行おうとすれば、ゲノム全体を一度にアセンブリングするといった、巨大な問題を解く必要が発生してくる。このためには現在のアルゴリズムはまだ力不足である。さらなるアルゴリズムの改良により、このような巨大なアセンブリングの問題も解けるように開発を行いたい。

5. 研究経費

平成１９年度見込：　5,000 千円

6. 平成19年度の研究成果

(1)知見・成果物・知的財産権等

・マウス 13 番染色体の配列の正確な決定

今回、難解読領域のモデルとして、第１３番染色体 65-68Mbp 付近の領域を解析した。解析開始時点では、この領域のデータベース（Ensemble 45 および NCBI Build36.1）上には、5 つの GAP（未解読領域）が存在していた(図 1A, 灰色の領域)。これらの GAP は、位置情報だけでなくその長さも未知であることから、この領域の塩基配列情報は極めて不正確である。この領域には遺伝子の重複が多く存在する事が示唆されており、その繰り返し配列が、断片化されて解読された塩基配列(BAC 配列)のアセンブリを困難にしていると考えられている。この第１３番染色体の未解読領域について、「高速相同性発見手法」を応用して、既読の BAC 配列の再アセンブリを行った。

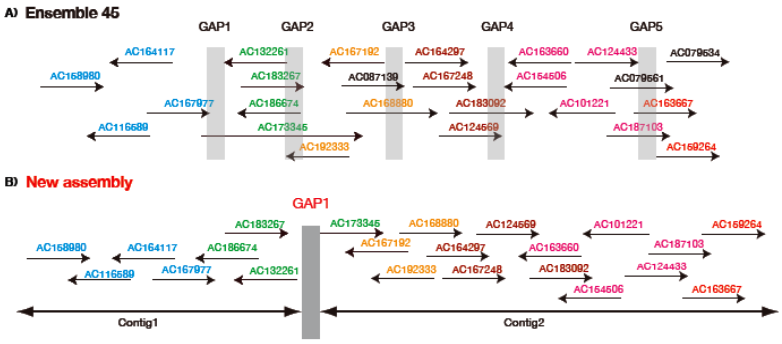

図1、第13番染色体65-68Mbp領域のBAC配列の並び。矢印はBAC配列を示している。
A) Database(Ensemble database ver. 45) 上のBAC配列の並び。
BAC配列が重なり合う領域に矛盾があるために、GAPが５つある。
B) 再アセンブリの結果新たに作成されたBAC配列の並び
BAC配列を正確に並べた結果、新たにContig1とContig2を得た。しかし、これらの２つの配列の間には、GAPが一カ所存在する。A)の黒色で示したBAC配列 (AC079561, AC079534, AC087139)は、他のBAC配列とアセンブリができなかったため、再アセンブリには用いられなかった。

再アセンブリの結果、５カ所あった GAP 領域を１つにすることができた（図 1B)。つまり、今まで不明であった、実際のギャップ領域の長さと位置の推定に成功した。これらの領域の多くは、サテライト配列等、数百塩基の単純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、GAP が生じていると考えられた。

また、今回新たに再構成された BAC アセンブリ配列は、NCBI で独立に再アセンブリが行われたと考えられる Build37.1 と非常に良く一致していた。つまり、今回用いたアセンブリ方法の正確性が実証された(図 2)。ただし、Build37.1 では GAP の無い一つの長い配列であるの対して、新たにアセンブリした配列では、GAP が一つ残っている。この理由は不明であるが、この領域近傍の BAC 配列自体が間違いであり、NCBI がこの領域近傍の BAC 配列を読み直した可能性が考えられる。もしくは、この GAP をまたぐような新たな BAC 配列が決定された可能性も考えられる。何れにせよ、この領域近傍のショットガンシークエンス配列を集め、今回用いたものと同様の方法でアセンブリを行うことにより、BAC 配列自体の妥当性を調べることができる。もしくは、この領域付近の配列を実験的に調べることにより、GAP 領域を埋めることができると考えている。

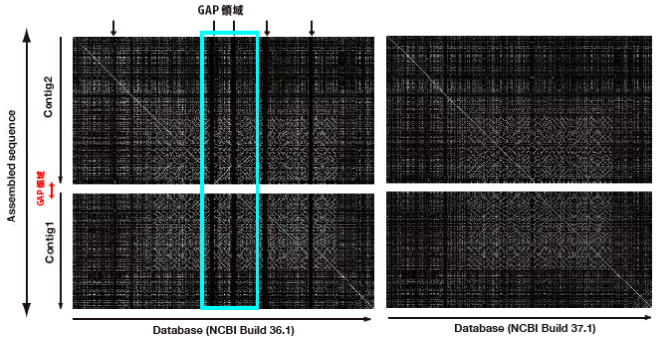

図2、 再アセンブリされた配列とデータベース (NCBIBuild36.1および37.1)との比較
NCBI build36.1とContig1とContig2の比較では、GAP付近の領域において、不一致が見られた（水色の線で囲った領域）。しかし、最近報告されたBuild37.1の配列と比較した結果、ほぼ全ての領域で一致していた。しかし、再アセンブリした配列にある１つのGAPが、Build37.1では存在しなかった。

今回の解析により、高速相同性検索法を用いたアセンブリ手法は、繰り返し配列の多い難解読領

域の塩基配列のアセンブリと、矛盾点（GAP）の同定に有用であることが示された。今回の手法の利点は、1)BAC 配列全体の比較ができるため、繰り返し配列と相同配列を視覚的に区別することが可能である、2)アセンブルできる BAC 配列のペアを選別することで、アセンブリの精度を高めることができる、3）アセンブリするペアの矛盾点を調べることによって、それぞれの BAC 配列の整合性、つまり BAC 配列のシークエンスアセンブリの精度を推定することができることである。

・正確な計算を行うことの優位性

今回の行った解析では､既存のアルゴリズムよりも高い精度で相同領域を発見するアルゴリズムを用いた。そのため、すでに決定されたマウス 13 番配列の中から、明らかにおかしなつなぎ方をしている部分を発見できたこと､および､より矛盾のないつなぎ方を見つけることに成功した。これは､より正確な相同領域発見アルゴリズムを用いることが､より矛盾のない正確な配列決定を行うために優位でありことがわかった。また同時に、すでに決定された配列の中にも、このような矛盾点が潜んでいる可能性をうかがわせることとなり､すでに決定した配列の検証作業が重要であろうとの知見を得るにいたった。

・アルゴリズムの実効性の検証

今回用いた相同領域検索アルゴリズムは､すでにその計算能力は確認されていたが､配列決定など実際の作業に用いたのは本プロジェクトが初めてであった。実際の作業を行う中で､必要となった機能を追加することで､アルゴリズムの実装をより使いやすくすることができ､また実用上のパフォーマンスについて評価を行うことができた。また､実用的に必要とされる速度と精度に関しても知見を得ることができ、これは今後の研究に生かされていくものである。

(2)成果発表等

＜論文発表＞

〔会議録〕

Takeaki Uno, “An Efficient Algorithm for Finding Similar Short Substrings from Large Scale String Data”, The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2008, to appear.

＜会議発表等＞

〔招待講演〕

・ 宇野 毅明, "大規模データ処理に対するアルゴリズム理論からのアプローチ", 回路とシステム軽井沢ワークショップ招待講演, 2007 年 4 月 13 日

〔一般講演〕

・ 宇野 毅明, “地球に優しいゲノム相同検索”, ミニシンポジウム:地球環境問題と計算限界, 2007 年 12 月 3 日

(3)その他の成果発表

・ Takeaki Uno, “Output Sensitive Algorithm for Finding Similar Objects”, Invited talk

on Symposium Combinatorial Days, Department of Computer Science, Swiss Federal Institute, 2/Jul/2007.

・ Juzoh Umemori, Ryouta Kondou, Takeaki Uno, Shigeki Yuasa, Tsuyoshi Koide, “Aberrant Neurological Development Caused by Genetic Incompatibility Between Two Wild-derived Strains, Oral presentation, 21st International Mammalian Genome Conference” Kyoto, 28/Oct-1/Nov/2007. （Journal Award : Science Award 受賞）

・ 宇野 毅明, “発見問題に対する列挙手法を用いた探索”, 生物情報解析研究センターセミナー 招待講演 2007 年 9 月 27 日

・ 宇野 毅明, 梅森 十三, 小出 剛“高速相同性発見手法を用いたゲノム解析とその応用”, 融合研究シンポジウム ポスター発表 2007 年 10 月 18 日

・ 梅森 十三, 小出 剛, 宇野 毅明. “高速相同性発見手法を用いたゲノム解析とその応用-新たなアセンブリ手法の開発-”, 口頭発表 若手研究クロストーク 2007 年 11 月 26-27 日

・ 近藤亮太, 梅森十三, 宇野毅明, 小出剛.“髄鞘形成異状をもたらす遺伝的不適合の原因遺伝子座 Genic1 の解析”, ポスター発表 第 30 回日本分子生物学会年会 2007 年 12 月 11-15 日