論 文 の 内 容 の 要 旨

論 文 題 目　　RAMEN: Genome Assembler
（ＲＡＭＥＮ：ゲノムアセンブラ）

氏　　名　　笠 原　雅 弘

ゲノム配列は遺伝学において最も基礎的な情報である。ゲノム配列はいまや、医学・生物学・農学などの幅広い分野でも様々な形で使われるようになった。しかし、その広範な応用にもかかわらず、ゲノム配列の決定は簡単ではなく、大量の実験と巧みなアルゴリズムを必要とする。最新のＤＮＡシークエンサーは一回の稼働で１キャピラリーあたり、連続する１０００塩基対を決定することができるが、例えば脊椎動物のような複雑なゲノムはそのサイズが数十億塩基対に達することもあり、ＤＮＡシークエンサーの読み取り長を遙かに超えている。そのため、ゲノムＤＮＡをランダムに裁断しＤＮＡシークエンサーで配列決定する「全ゲノムショットガン法」がゲノム配列を決定するために使われている。

しかし当然のことながら、ランダムなゲノムＤＮＡ断片配列はゲノム上での位置が分からない。そこで、ランダムＤＮＡ断片配列の集合から元のゲノム配列を再構成するコンピューターアルゴリズムが必要となる。この操作を行うコンピュータープログラムを「ゲノムアセンブラ」と呼び、世界中で多くの研究グループがこの問題に取り組み数百篇にも及ぶ数の論文を発表している。しかし、大きなゲノム配列に対する配列解読プロジェクトにおいて見られるアセンブリ上の問題は、今まで詳細かつ実践的には充分に語られてこなかった。これはおそらく、扱うデータが大きすぎるために、問題を発見し記述すること自体が難しかったからであろう。

膨大なデータを扱う際には、計算資源の制約から、単純なアルゴリズムやデータ構造しか使うことはできない。また、ゲノムアセンブリの性質は、解読対象の種やシークエンシ

ングセンターに応じて大きく異なることもある。ある状況では有効であった改良が、種やシークエンシングセンターが異なると改悪である場合すらある。このような理由で、大きなゲノムに対するアセンブリアルゴリズムは体系的に論じられたことは無かった。本稿ではまず、大きなゲノム配列を解読するための既存のアセンブリアルゴリズムを体系的に紹介する。ゲノムアセンブリにおける問題点の多くは、ＤＮＡシークエンサーの技術的限界に由来しているため、ＤＮＡシークエンシング技術も概説する。

次に、本稿では、私が新たに開発したゲノムアセンブラであるＲＡＭＥＮを、ステップ毎の詳細なアルゴリズムとともに報告する。ＲＡＭＥＮはメダカゲノム解読を目指して開発された。アルゴリズムの詳細に加えて、大きなゲノム配列解読において遭遇する問題点を実際のゲノムプロジェクトを例に報告する。具体的には、ベクターマスク問題や多型、系統的シークエンシングエラー、互い違いのスキャッフォルド、クローンサイズ推定の精度不足などについて論じる。これらの問題点に対処するアルゴリズムと、その背景となる考え方についても同時に述べる。

また、ＲＡＭＥＮの性能を測るために、メダカ（*Ｏｒｙｚｉａｓ　ｌａｔｉｐｅｓ*）ゲノムをアセンブルした。ショットガンリードの配列カバー率は１０倍以上であり、総計クローンカバー率５３．７倍分のプラスミド・フォスミド・ＢＡＣ等のメイトペア情報が与えられた。アセンブリの結果、Ｎ５０スキャッフォルド長は１．４１Ｍｂに達した。また、メダカの遺伝学的地図を統合し、ウルトラコンティグ長は５．１Ｍｂに達し、約９０％の塩基が染色体と関連付けられた。アセンブリと完成ＢＡＣ配列との比較から、コンティグの端１００ｂｐを除くと塩基精度は９９．９６％であると推定された。

これらの結果から、ＲＡＭＥＮは脊椎動物のゲノムをアセンブルするのに十分な精度とスケーラビリティを備えていることが示唆された。

また、本稿の最後に、大量のＤＮＡシークエンサーから産出される膨大なデータを処理するためにはアルゴリズムを実装するうえで特別な注意が必要であったという教訓を紹介する。本稿では並列プログラミングのフレームワークについて注目して議論するが、これは次世代シークエンサー用のゲノムアセンブラを実装する際には特に重要となるだろう。