審 査 の 結 果 の 要 旨

氏　　名　　舛谷　万象

ゲノム配列の解読手法が提案されて約半世紀間、ゲノム配列は現代生命科学の中心的な情報として役立ってきている。しかし、「解読完了」を宣言されたゲノム配列の数々とは裏腹に、ゲノム配列を完全に解読することは技術的には容易でなく、ヒトを含め様々な生物のゲノム配列の多くは解読完了とされているのにもかかわらず、実際にはその配列は不完全な「概要状態」のまま長い間を過ごしてきた。様々なゲノム配列の完全解読へ向けては幾つかの技術的な進歩が必要出会った。例えば、商用ＤＮＡシークエンサーの読み取り塩基精度は徐々に改善し、また、解読可能な DNA 断片も大きく伸長してきている。特に 2015 年以降は、１万～100 万塩基対の DNA 断片を効率的に解読可能な技術（ロングリード・シーケンシング）が普及し、反復配列を正しくアセンブリすることが難しい短い DNA 断片では無く、長い DNA 断片をアセンブルしてゲノム全体を完全に復元することを期待できる状況となった。

その一方で、既存の技術では完全な復元が非常に困難な領域も明らかになってきた。ゲノム配列の復元が技術的に難しかった例として、ヒトゲノムでは個人間での違いが大きく多様であることが知られている主要組織適合遺伝子複合体領域（major histocompatibility complex; MHC）、ヘテロプラズミーの影響もあり構造上の多様度が大きいことが知られている植物のミトコンドリア・ゲノムがある。両者共に、ゲノム配列の復元が技術的に難しかった主たる原因は、長い配列が重複して存在しており、重複配列由来の DNA 断片配列を上手く識別して分類するのが難しいことにあった。この分類問題を解くために、チャンクと呼ぶ数千塩基の領域にゲノム配列を分割し、チャンク上で重複配列由来の DNA 断片を適切に分類できないか、というアイデアを発想し、そのアイデアの実用性を舛谷万象は検証した。特に、短いＤＮＡ断片とは異なり、長いＤＮＡ断片を読み取る際にはシークエンシングエラーの割合が比較的高いため、そのような読み取りエラーの存在下であっても頑健に重複配列由来の配列を見分ける必要があった。

それ以前の手法との着想の違いは、適度に長いチャンクを利用すれば、２つの重複した領域に特異的な塩基変化を捉えやすいことに着目した点にある。しかし、シークエンシングエラーの存在下でこのような分類を頑健に精度良く行うためにはエラーと本来の塩基多型を見分ける工夫が必要である。本研究では塩基置換・挿入・削除操作、あるいは複数領域間の塩基多型発生のそれぞれについ

て尤度を表現する行列を設計し、塩基置換・挿入・削除などの操作に対して尤度が低くならない、即ち尤度を局所最大化する DNA 配列のクラスター（コンセンサス配列）を計算するアルゴリズムを新しく設計しゲノムアセンブリへと応用しており、本論文の特筆すべき成果となっている。この結果、ゲノム上で対応している配列を高精度でクラスタリングできることを、様々な正解が既知のベンチマークデータや、未知の実データから実証している。

また、植物ミトコンドリアゲノムのゲノムアセンブリにおいては、アセンブリした配列に対する断片ＤＮＡ配列のアラインメントデータから、２番目以降に存在比が高いミトコンドリアのゲノム形態を明らかにするなど、ミトコンドリアゲノムの構造的な多様性をロングリードを用いた系統的な手法によりめて明らかにした。

今後のゲノムアセンブリ戦略の策定に有用な示唆も得ている。例えば、DNA 断片配列を読み取る装置について受け付けうる断片配列超と読み取り精度の間にトレードオフがあるとすると、長いアセンブリ結果を得るには断片ＤＮＡの読み取り精度が多少低くとも、長い DNA 断片を用いる方が総合的には有利であることを実験的に示している。また、多様性が大きな領域は、標準配列をガイドとして用いるアセンブリ方式より、本論文が提案するアルゴリズムを用いてゼロから（*de novo*）アセンブリするのが有効であることも例示している。

なお，本論文は，有村慎一氏，森下真一氏との共同研究であるが，論文提出者が主体となって解析及び検証を行なったもので，論文提出者の寄与が十分であると判断する。

よって本論文は博士（科学）の学位請求論文として合格と認められる。

以上１７７５字