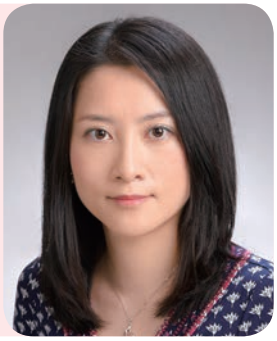
大規模ゲノムデータの解析技術の開発



早稲田大学　基幹理工学部　教授
清水　佳奈
〔お問い合わせ先〕 TEL：03-5286-3344　E-MAIL：shimizu.kana@waseda.jp

研究の背景

近年、ゲノム配列を決定する装置（シークエンサー）の性能が飛躍的に高まったことにより、シークエンサーから得られた膨大なデータを効率よく解析する手法の開発が望まれています。シークエンサーはゲノム配列を端から端まで一度に読み取るのではなく、数百塩基ほどの長さをバラバラに読み取ります。そのため、さまざまな情報解析を行うには、断片配列同士を比較して類似する配列を発見する情報処理が必要です。

研究の成果

従来の手法では、類似配列を発見するためにN本の配列に対してNの二乗に比例する計算量が必要でした。これを改善するため、まず、私たちは類似する配列同士に共通する部分配列の法則を明らかにしました。そして、その法則をうまく利用することによって正解の候補を効率的に絞り込み、Nに比例する計算量を実現するアルゴリズムの設計に成功しました。実験では、１万～１千万本の断片配列に対する計算が、従来の手法と比較して数十～数千倍高速化されたことが確かめられました（図１）。また、断片配列同士の類似度から分類解析を行うソフトウェアの実装も行いました（図２）。

開発したアルゴリズムはがんゲノム解析にも応用できます。がん細胞ではゲノムの特定の領域が入れ替わるリアレンジメントと呼ばれる現象が生じることがありま す。従来の手法では、シークエンサーから得られた断片配列を参照ゲノム配列と呼ばれる標準的なゲノム配列と比較してリアレンジメントが起きている場所を特定します。しかし、個体特有の配列が断片配列に含まれている場合は解析がうまくいかないという問題がありました。これに対して、私たちは、上記の高速アルゴリズムを用いて、同一個体の正常細胞から得られたデータと、がん細胞から得られたデータを直接比較する手法を開発しました。この手法を用いると、参照ゲノムの偏りによる影響を受けにくい解析を実現することができます。

今後の展望

今回開発した手法は、さまざまなゲノムデータの分類解析や構造変異解析などに応用することが可能です。今後は、腸内細菌などのメタゲノム解析やより複雑ながんゲノム解析、さらに多数の個体の情報を組み込んだ参照ゲノムグラフの構築に応用していきたいです。また、ゲノムシークエンサーの性能は日進月歩で向上しており、より効率のよい情報解析技術が求められています。そのため、これまでに開発してきたアルゴリズムの性能をさらに高める研究も必要であると考えています。

関連する科研費

2010−2011年度　若手研究（B）「ギガシークエンスデータの高速解析技術の開発」
2014−2016年度　挑戦的萌芽研究「類似ゲノムの差異を逃さないDe novoゲノム解析技術の開発」

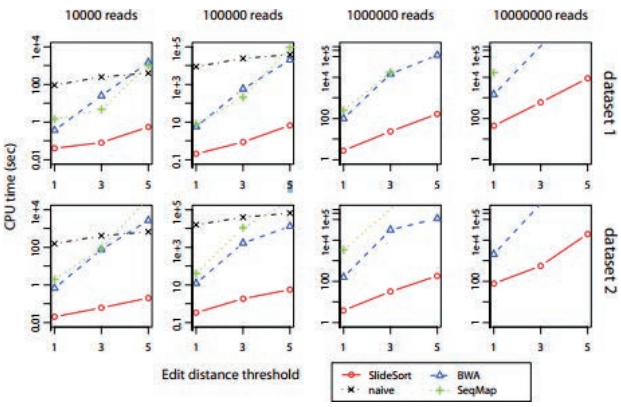


図１　従来手法との性能比較（赤線が開発したアルゴリズム）

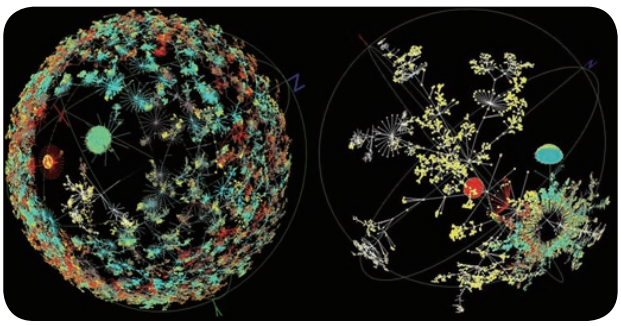


図２　断片配列の分類結果の例