論文内容の要旨

論文題目　Construction of systems for large-scale comparative analyses of microbial genomes
大規模な微生物ゲノム比較解析のためのシステムの構築

氏　　名　　内　山　郁　夫

決定されたゲノム配列の数は急速な増加を続けており、中でも微生物ではすでに数百におよぶゲノム配列が決定されている。こうした中で、大量のゲノムデータから比較解析を通じて有用な生物学的知見を引き出すことが、ゲノム研究の重要な目標のひとつとなっている。特に近年、互いに近縁なゲノムの配列データが急速に蓄積するようになり、これらの比較解析も可能になってきた。様々な進化距離にあるゲノムを比較することにより、生物学的な機能や進化プロセスに関する、様々な種類の情報を抽出することが可能になる。このため、大規模な比較ゲノム研究を進める上で、遠縁ゲノム比較と近縁ゲノム比較の特長を活かしつつ、これらを並行して進めることが重要である。こうしたことを念頭に置いて、本研究では、主として原核生物（真正細菌と古細菌）を対象として、大規模な比較ゲノム研究のための基礎的な手法やツールの開発を行った。原核生物については、近年の大量のゲノム配列決定を背景として、これまで未知であった自然界における多様性の実態やその進化プロセスの解明が、急速に進むことが期待されている。

比較ゲノム研究を行う上で、ゲノム間でオーソログの対応付けを行うことは最も基礎的な工程であり、この対応付けに基づいて様々な比較解析を行うことが可能になる。従来、２ゲノム間のオーソログの対応付けは、多くの場合に「双方向ベストヒット」(Bidirectional Best Hit, BBH)という基準を用いて行われており、多ゲノム間で比較する場合は、これに何らかのクラスタリングアルゴリズムを組み合わせることが多かった。しかし、この手法はいくつかの問題点を抱えているため、より信頼できる結果を得るためには Clusters of Orthologous Groups (COGs) データベースのような、専門家による手作業を交えて分類されたデータベースを利用する必要があった。しかしながら、こうしたアプローチを大規模なデータに適用するには大きな困難が伴う。そこで本研究では、オーソログの対応付けを自動的に精度良く行う手法 DomClust を開発し、それに基づ

いて微生物ゲノム間の比較解析を行うためのシステム MBGD を構築した。

MBGD は微生物ゲノムの比較解析ワークベンチであり、その中心機能はオーソログ対応表の自動的な作成と、それを活用したゲノム比較解析機能にある。このため、複数のゲノム間であらかじめ遺伝子の総当たりホモロジー検索を行い、その結果をデータベースに格納している。MBGD では、このデータを用いてオーソログの対応付けを自動的に行う機能が実装されているため、最新のデータを含むオーソログ表を迅速に提供できるだけでなく、利用者が指定した生物種集合に対するオーソログ表を動的に作成することも可能である。後者は、特定の系統群に属する類縁ゲノム間で比較を行いたい場合に特に有用な機能である。作成したオーソログ表はデータベースに格納され、利用者は、これに基づいてオーソログの各ゲノム内での有無に基づく「系統パターン解析」や、オーソログ周辺の並び順の比較、各オーソログの詳細な系統解析などを行うことができる。本システムを使用した解析例として、系統パターン等によって機能的に関連する遺伝子集合（機能モジュール）を見つける問題について論じた。

DomClust は、MBGD で用いられている、ゲノム間のオーソログ対応付けを行うプログラムである。このプログラムは、総当たりのホモロジー検索結果を用いて、標準的な階層的クラスタリングアルゴリズムである UPGMA に基づいて分類を行うが、クラスタリングの途中でドメイン融合や分裂を検出し、必要に応じてドメインの分割を行う処理が組み込まれている。その後、構築された階層的ツリーを分断しながら、種内パラログを含まないようにグループを分割することによって、オーソロガスなグループを作成する。このようにして、この手法はオーソログ分類を行う上で、最小限のドメイン分割を行うことができる。この手法を評価するため、COG データベースを参照データベースとして、それとの一致度で分類を評価する手続きを作成した。比較のため、従来行われているような BBH 基準に各種のクラスタリングアルゴリズムを組み合わせて分類する手法についても同様に評価した。その結果、BBH 基準に基づく分類と比較すると、DomClust による分類は COG とよりよく一致することが示された。さらに、データセットが増える際の影響を調べるため、リリース時期の異なるデータセット間でクラスタリング結果を比較することにより、クラスタリングの安定性を評価した。その結果、我々の手法は BBH に基づく手法と比べて比較的よい安定性を示すことがわかった。

一方、より近縁のゲノムを比較する場合には、ゲノム間アライメントを用いることが多い。ゲノム間アライメントは、主に遺伝子のコード領域や制御領域などの保存領域を見つける目的で広く用いられており、すでに多数の手法やツールが開発されている。しかしながら、ゲノムの進化過程を考察するためには、挿入、欠失、逆位、重複を含む、ゲノム構造の変化を解析する必要があるが、こうした目的に適したツールは多くない。

そこで、そうしたゲノム構造変化の観察に適したツールとして CGAT を開発した。CGAT は、クライアント・サーバ型のソフトウェアで、アライメントの計算やデータの管理を行うサーバと、アライメントの表示を行うクライアントからなる。クライアントは、ドットプロット表示とアライメント表示の組み合わせによって、大域的な構造変化から局所的な構造変化までの詳細な観察を容易にするツールである。それぞれがズームの機能を持っており、オーソロガスなアライメントに沿って表示域の移動ができる。また、各ゲノム上に種々の情報を表示して比較する機能を持っており、特に様々な種類の繰り返し構造や、挿入因子などの可動因子の位置を表示することによって、それらとゲノム多型との関連を調べることができる。一方、サーバは、既存の局所アライメントプログラムを用いて大規模なゲノム間アライメントを計算するための共通のプロトコルを実装している。これには、問い合わせ配列を分割してアライメントを計算し、つなぎ合わせる処理や、隣接したアライメントを連結したスコアを計算し、オーソログを同定する処理などが含まれる。この機能によって CGAT は、同じゲノム対のアライメントを複数のプログラムで行って結果を比較し、特定のヒューリスティクスに起因したアライメントの誤りを検出してこれを回避する、といった目的に使うこともできる。具体的な解析例として、ヘリコバクター・ピロリ菌２株の比較において、ゲノム多型に関連する興味深い構造を表示する例をいくつか示し、さらに複数のアライメントプログラムの比較について詳しく論じた。

最後に、近縁から遠縁まで様々な類縁度のゲノムを比較する問題について考察し、将来展望として、それらを統合した比較ゲノム解析を行うためのシステムについて論じた。