論 文 の 内 容 の 要 旨

論文題目　ゲノム情報解析の再現性と再利用性を向上させる情報基盤の設計
氏　　名　山中　遼太

次世代シークエンサーによりゲノム配列を解読するコストが下がり，ゲノム配列の同定に留まらず，配列同士を比較して得られるバリエーションや変異の情報，mRNA や miRNA の定量による遺伝子の発現量，ヒストン修飾や DNA のメチル化状態といったエピゲノム情報，など多様なデータを産出することができるようになった. 生命科学においてこれら大量のデータから仮説を構築するアプローチが注目されているとともに, 医療において臨床シークエンシングと呼ばれる診断への応用も期待されており，信頼性の高いゲノム情報の取得と収集が求められている.

ゲノム情報解析には大きく分けて，データ処理とデータ統合の二つの解析フェーズがある．まず，次世代シークエンサーから出力されるデータは DNA 配列または RNA 配列の断片であり，実験結果として解釈できる情報を得るためには計算機を用いてこれら配列断片データを実験系に応じて処理する必要がある．次に，こうして得られたデータをデータベースに集積し統合することで，ある条件に適合するデータを検索する他，群の分類や比較といったデータ・マイニングが可能になる.

ここで，それぞれの解析フェーズにおいて，解析結果の再現性が低いことや解析手順が共有されていないといった課題がある．データ処理のフェーズにおいては，処理の一連の手順（ワークフロー）と実行環境が研究機関単位で管理されており、この維持と共有ができないことで再現性が損なわれ得る．また，データ統合のフェーズにおいては，複数のデータベースのデータの統合がそれぞれのアプリケーション開発者によって実施されているため, データ統合のプロセスが再利用できず冗長な作業が発生すると同時に統合プロセスの再現性が低い．そこで，本研究では，データ処理とデータ統合のそれぞれの解析フェーズに対して, 再現性と再利用性を向上させるための情報基盤を設計し，これを構築および検証している.

第一に，データ処理フェーズの再現性と再利用性の向上のため，複数の研究機関のデータ処理ワークフローを実行可能な解析環境を仮想マシン上に構築した. この仮想マシンを使用することで, 過去に実行したワークフローを確実に再現できると共に, 異なる研究機関でワークフローを共有することができる．また，ワークフローはゲノム情報解析で広く用いられているワークフロー管理システム「Galaxy」上で管理されており，ウェブのインターフェイスを通して情報系研究者以外でも操作することが可能である．環境の構築手順はコード化されており、ワークフローの追加や更新に応じて，この仮想マシンを定期的に更新することができる。

この結果，異なる研究機関の複数の開発者が作成したワークフローが仮想マシン上で実行可能となり、2015 年 4 月のワークショップ「Galaxy Workshop Tokyo 2015」においておよそ 100 人のユーザーに向けて発表された。この環境には，ChIP-seq, RNA-seq, Bisulfite-seq, Variant Calling といった様々な実験系に対するワークフローが収集されており，それぞれのワークフローについての解説が同一のウェブ・サイト上に掲載されている．この環境は PC、サーバー・クラスタ、クラウドといった異なる計算基盤上で動作し、各ワークフローを使用した解析結果が同一であることを試験によって示している。今後，ワークフローを継続的に更新するとともに，実際にデータ処理の再現性と再利用性が向上していることを評価するためにはこの解析環境の利用状況を追跡していく必要があるだろう．

第二に，ゲノム情報のデータ統合の再現性と再利用性の向上のため，セマンティック・ウェブ技術を用いたデータ統合とアプリケーション開発を，公共がんゲノム･データを用いて検証した．国際がんゲノム・コンソーシアム（ICGC）が公開しているがんにおける塩基変異データから RDF データを生成し，遺伝子名やパスウェイに関する外部のデータベース由来の RDF データと統合した．さらに，この統合済みデータを用いてがんの症例と塩基変異の情報を検索できるデータ・ポータルをセマンティック・ウェブ・アプリケーションとして実装した．

この結果，ICGC データから生成された RDF データ、および外部データとの統合用 RDF データが公開された．このデータの利用者は既に統合されたデータセットを入手することで，データ統合の工数をかけずに，データベース横断的な検索のためのクエリを作成することができる．また，各アプリケーション開発者がデータ提供者の意図したデータ統合を使用することにより，クエリ結果の再現性も向上できる．さらに、このような RDF データを使用した実用的なアプリケーションが実装できることを示した．

本研究では，データ駆動型のゲノム科学研究の発展，およびゲノム情報解析の医療への応用のために，解析の再現性と再利用性を大幅に向上させる必要があることを指摘し，そのためのデータ処理とデータ統合の情報基盤の設計をそれぞれ議論した。その上で，既存の仮想化技術やセマンティック・ウェブ技術を用いて、実際にユーザーが活用できる解析環境や技術検証のためのウェブ・アプリケーションを作成した．このような設計例を基に，今後もゲノム情報解析の再現性と再利用性を高める努力が継続されることで、信頼できる情報基盤が整備され、ゲノム情報の医療への応用が加速することが期待できる。