審 査 の 結 果 の 要 旨

氏　　名　　山中　遼太

本論文「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」では、次世代シークエンサーにより得られるゲノム情報の解析には高い再現性と再利用性が求められていることを提示し、これを実現するための情報基盤の設計を議論している。このために、ゲノム情報解析をデータ処理とデータ統合の二つのフェーズに分けて、それぞれのフェーズについて現状の問題点とそれに対する解決方法を提示し、検証している。以下では、これらそれぞれの研究課題について、審査における評価を記載する。

第一に、ゲノム情報のデータ処理のフェーズについて、申請者は次世代シークエンサーから出力された大量の配列断片から解釈可能な情報を抽出するための処理ワークフローが、異なる計算環境の間で再現可能ではないことや、異なる研究室の間で再利用されていないことを問題点として提示した。これを解決するため、申請者はワークフロー管理システムと仮想化技術を組み合わせて、複数の研究者で同一の計算環境を作成して、この計算環境を仮想マシンとして配布し、共有するという枠組みを提案した。

評価できる点として、実際にこの枠組みが立ち上げられたことが挙げられた。複数の研究者による開発者会議が継続的に実施され、提供された 10 程度のワークフローを含む計算環境が仮想マシンとして配布されており、ワークショップなどを通して既に多くの利用者がこの計算環境を動作させている。さらに、申請者は、この計算環境の潜在的な利用者に情報提供するためのワークショップを開催して 100 人近くの参加者を集め、そのフィードバックを得るなど、長期的な視野に立ってこの計算環境を広める活動を実施しており、海外コミュニティの活動とも連携している。

疑問点の一つ目として、仮想マシンを使用した環境のパッケージ化とその配布は既存の技術を用いた方法であり、技術的な新規性があるとはいえないこと

が挙げられた。これに対し、申請者は同一の計算環境を共有する方法として、仮想マシン以外にも環境構築手順のコード化やコンテナ技術といった方法を検討および検証しており、その上で、現時点で最適な方法として仮想マシンを選択している。このため、申請者は技術的な新規性を求めるよりも計算環境の共有という目的に対する実用性を優先したといえる。

疑問点の二つ目として、配布している環境に含まれる解析ワークフローは 10 程度とまだ少なく、実際の利用者数や利用回数といった情報もないため、提案した枠組みによって再現性や再利用性が向上したかどうか定量的な裏付けができていないことが挙げられた。これに対し、申請者は、再現性と再利用性の評価指標のうち、利用状況や論文への引用といった指標は一定の利用者が得られた後の将来課題としており、現時点では異なる計算機上でワークフローの実行結果が同一であるという検証結果のみを指標としている。

第二に、ゲノム情報のデータ統合のフェーズについて、申請者は、複数のデータベースのデータ統合がアプリケーションごとに実施されていることで、データ統合の再現性と再利用性が低いことを問題点として提示した。そこで、これを解決するためにセマンティック・ウェブ技術のひとつである RDF を用いたデータ統合を検証した。大量のデータが収集されている国際がんゲノム・コンソーシアムのデータから RDF データを生成し、これを公開すると共に、この RDF データと外部の RDF データを統合したデータ・セットを使用してデータ・ポータルを実装した。

評価できる点として、がんゲノム RDF データとデータ・ポータル開発手法の新規性が挙げられた。多くの生命科学データから RDF データが生成されているが、がんゲノム・データの RDF データは今までに公開されておらず、この RDF スキーマとデータの作成は新規性があるといえる。また、生命科学分野において RDF データを使用した実際のアプリケーションは少ないため、今回のデータ・ポータルの実装のために使われた手法やノウハウにも新規性が認められる。

疑問点の一つ目として、新しく生成した RDF データが再利用されていないことが挙げられた。データ・ポータルの実装において外部の RDF データの再利用を実践している一方、新しく生成した RDF データが再利用されるためには、これががんゲノムの RDF データとして広く認知される必要があり、現時点ではまだ再利用されていない。これに対し、申請者はこの RDF データについて海外の学会で発表している他、RDF データのポータル・サイトへの登録や論文投稿の

準備を進めている。このため、今後、このデータが再利用されると期待できる。

疑問点の二つ目として、新規に実装したデータ・ポータルに実用上の優位点がないことが挙げられた。ここで実装したデータ・ポータルは、RDF データを使って既存のデータ・ポータルと同等のものが実装できることを検証するためのものであり、性能と機能の面で既存のデータ・ポータルより長けている点がない。これに対して、申請者は、性能の向上のためにキャッシング機能などを開発している一方、根本的な性能の問題はシステム設計とデータベース・ソフトウェアの将来課題としている。また、機能については、統合後のデータで容易に可能となる検索の例を示すことで、新規性のあるアプリケーションを作成できることを示唆している。

以上の通り、審査において本論文の研究内容が評価された。それと同時にいくつかの疑問点が取り上げられたが、これらの点に対しても論文中において、適切な対応、議論の展開、将来展望の提示がなされていることが示された。

よって本論文は博士（工学）の学位請求論文として合格と認められる。