【24】

氏名	瀬尾淳哉
博士の専攻分野の名称	博士（情報科学）
学位記番号	第　２３９２８　号
学位授与年月日	平成22年３月23日
学位授与の要件	学位規則第４条第１項該当 情報科学研究科バイオ情報工学専攻
学位論文名	生物情報解析におけるワークフローの検索手法と実行方式に関する研究
論文審査委員	（主査） 教授　松田　秀雄 （副査） 教授　清水　浩　教授　四方　哲也　教授　前田　太郎

論文内容の要旨

本研究では生物情報の分野におけるワークフローの検索手法と実行方式に関する研究を行った．

まず作成支援のためのワークフロー検索手法の提案を行った．生物情報解析では複数の解析ツールやデータベースを組み合わせて用いることが一般的であるが，同種の解析ツールが多数存在するため，それらの選択や組み合わせを決定することが困難である．そのため，レポジトリに蓄積された既存のワークフローを新規ワークフローの作成のために再利用することが行われている．しかし，ワークフローの解析内容を反映した検索方法が存在せず，再利用のためのワークフローを検索することが困難であった．この問題に対し本研究では，ワークフロー中の単語から特徴を表すものを抽出し，その単語に対する検索処理を追加することにより，解析内容を反映したワークフロー検索手法を提案した．実際のレポジトリのワークフローを対象に検索実験を行い，検索結果の精度の計測を行った．それにより，探したい解析処理を含むワークフローを高精度で検索できることを示した．

次にワークフローを効率的に実行するための方式に関する研究を行った．生物情報解析ワークフローには，個々の解析ツールの処理時間に比べてツール間で受け渡されるデータの転送時間の占める割合が大きい，ツールによる解析処理以外にフォーマット変換などのデータ処理や条件分岐などの実行制御が必要といった特徴がある．そのようなワークフローを既存の実行方式で実行した場合，ワークフローの実行制御を行っている実行エンジンへとデータが集中してしまい，その実行エンジンを介したデータ転送がボトルネックとなってしまうことが問題となっていた．そこで本研究では，解析ツールの結果に対する処理を実行エンジン外で行うことを可能とし，データ転送を効率的に行うことが可能となる実行方式を提案した．提案方式を用いることにより，任意のワークフローにおいて，実行エンジン側へとデータが戻されることなく効率的に実行することが

ローにおいて，実行エンジン側へとデータが戻されることなく効率的に実行することが可能となる．実際のワークフローの実行を想定した実験環境において，提案方式と既存の実行方式での実行時間を測定し，これをもとに各方式間の性能比較を行うことで提案方式の有効性を確認した．以上のことから，提案方式により生物情報解析ワークフローを効率的に実行できることを示した．

論文審査の結果の要旨

本論文は、生物情報解析の分野におけるワークフローの検索手法と実行方式について述べている。生物情報解析では、外部のデータベースや解析ツールを利用する一連の処理手順をワークフローで表現して実行することが広く行われている。本論文では、ワークフローの作成時に、既存のワークフローを部品として再利用するための類似ワークフローの検索手法と、分散処理環境でのワークフローの効率的な実行方式が提案されている。

まず、ワークフローの作成において、そのワークフローで表現すべき解析内容に類似した既存のワークフローを、ワークフローのレポジトリから検索する手法について述べている。本手法により、過去に作成されたワークフローの中から、所望の解析内容に類似した解析を行っているものを検索し再利用することで、新たなワークフローの作成が効率的に行えることが示されている。情報検索の分野で使われている、文書中の出現頻度で単語を重み付けする手法であるTF-IDF法を応用したワークフロー検索手法を考案することで、高精度な検索を達成している。

次に、分散処理環境での生物情報解析のワークフローの実行時に、サーバ間のデータ転送量を削減することにより、効率的にワークフローを実行する方式について述べている。生物情報解析では、解析に必要なデータベースや解析ツールが多種多様であり入出力データの形式が統一されていないため、ワークフローで記述された一連の解析処理の実行過程で、ある解析処理と次の解析処理の間に条件分岐やデータ加工といった中間データ処理がしばしば必要となることが指摘されている。従来のワークフロー実行方式では、このような中間データ処理の実行では解析処理サーバからワークフローの実行を制御する実行エンジンへのデータ転送が必要となるが、このデータ転送時間がワークフロー実行時間に対して大きな割合を占めることが示されている。中間データ処理を実行エンジンではなく、解析処理を行うサーバ側で処理する実行方式を考案することにより、時間のかかるデータ転送を削減し、ワークフロー実行時間の短縮を達成している。

これらの研究成果は、計算機を用いた生物情報解析における解析処理の効率化を推進し、生物情報科学の研究の発展に大きく貢献するものである。

よって、博士（情報科学）の学位論文として価値のあるものと認める。