【1】

氏名	木戸善之（きど よしゆき）
博士の専攻分野の名称	博士（情報科学）
学位記番号	第　２２３７５　号
学位授与年月日	平成20年６月13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	グリッドを用いた分散環境におけるバイオインフォマティクスツール実行支援に関する研究
論文審査委員	（主査） 教授　松田　秀雄 （副査） 教授　清水　浩　教授　前田　太郎　教授　四方　哲也

論文内容の要旨

本研究ではライフサイエンス研究領域で，研究組織間で協力体制を築く手段の1つとしてグリッドコンピューティングを用いた共同利用システムを提案した．グリッドコンピューティングでは，複数の組織間を統合し，利用できる環境を構築することで，従来の組織の壁を越え仮想組織の形成を促し，共同研究の加速化を目指すものである．本研究では様々な組織の研究者が利用する共同利用施設におけるデータの機密性確保について取り組み，GSI-SFSを利用し，ファイルパス名の隠蔽化した．それにより他のユーザにはどのようなファイルを利用しているかを知られないシステムを開発し，このシステムによりユーザは他の組織にどのような研究をしているかを知られることなく，複数の組織間にまたがる計算機クラスタを利用したデータ解析やゲノムの相同性検索が可能となった．

次に増加し続けるデータに対応するため，データ取得方法について研究を行った．この研究の背景としては，爆発的に増加するゲノムデータの問題が挙げられる．バイオインフォマティクスでのゲノム解析手法では，多くの生物種のゲノムデータとの比較解析が求められており，研究者が利用するデータ量は増加する一方である．このようなゲノム解析の研究では，ゲノムデータベースをローカル環境に展開し，独自のゲノム解析環境を構築する必要がある．ゲノムデータベースを公開しているデータベースサイトでは解析のためのWebインタフェースやWebサービスも公開しているが，これらは公開元が定義したデータベースと解析サービスのみしか利用できない．つまり特定のデータセット（例えば特定の生物種）のみに限定した解析を行いたい場合，Webインタフェースでは相同性検索の結果に対して手動で絞込みを行う必要があり手間がかかってしまう．よって研究者らは独自の解析環境を用意する必要に迫られる．従来では独自の解析環境のためのデータベースの取得方法としてミラーリングが行われてきた．ミラーリングはデータベースをローカル環境に複製を作る技術であり，データベース全体の複製をとるが，ゲノムデータが爆発的に増加している現状で完全な複製を維持するのは限界がある．以上の問題を解決するため，ゲノム解析環境においてデータステージングの適用を提案した．データステージングは，解析などのツール実行単位であるジョブの開始直前と終了直後に必要なデータファイルを計算機に転送する技術である．データステージングをゲノム解析環境に適用することで，必要なデータを必要なときに取得することができ，ローカル環境の計算機のストレージを圧迫することなく解析が可能となった．実験ではデータステージングで構築した解析環境とミラーリングによって作成された解析環境を比較することで，提案手法の有効性を評価した．データステージングによるファイル転送のオーバヘッドが計測されたが，解析環境で利用する計算ノードを増やすことで，ゲノムデータの比較解析を効率的に行うことができることを示した．

論文審査の結果の要旨

本研究ではライフサイエンス研究において共同利用計算機を安全に利用するためのファイルパス名隠蔽化手法を提案している。共同利用計算機は他者が利用するため、ファイルパス名を隠蔽化する技術が必要とされており、またライフサイエンス研究で利用するデータファイル数は増加していると述べている。従来手法であるメタデータを利用したファイルパス名隠蔽化では、メタデータに実ファイルパス名の写像として暗号化したファイルパス名を持たせることで、ファイル名を隠蔽化することができるが、ファイルの更新が増加した場合、メタデータの更新も発生するためオーバヘッドが増加すると本研究では指摘している。そこで本研究ではGSI-SFSを用いたファイルパス名隠蔽化手法を提案している。提案手法ではデータファイルを置くサーバに対し共同利用計算機から直接マウントするときに、他のユーザからマウントポイント自体を隠蔽することでファイルパス名の隠蔽化を実現している。またメタデータを利用していないことから、従来手法よりもファイル更新時におけるオーバヘッドが軽減できることを示している。

次に増加するゲノムデータの問題をあげている。ライフサイエンス研究では多くの生物種の完全ゲノムと比較解析が求められており、研究者が利用するデータ量は増加する一方であることを指摘している。完全ゲノムとの比較解析を公共のWebサービスで行うには、解析の実行と解析結果の送信が直列で行われるため、解析結果の送信を制御することができず、非効率であると本研究では指摘している。一方、データベース全体をコピーするミラーリングを用いてユーザの独自環境でゲノム解析を行う場合には、コピーのためのデータ転送量が問題となると本研究では指摘している。これらの問題点を解決するためにデータステージングを用いたゲノム解析を提案している。データステージングは解析を実行する直前に解析に必要なデータのみを転送し、解析終了直後に結果データを転送する手法である。提案手法ではデータ転送と解析実行を非同期で行い、データ転送と解析実行を同時並行することで、計算ノードを増加させた場合、解析時間の総和は短くなることを、Webサービスとの比較実験で示している。また計算ノード毎にデータベースの更新が発生するミラーリングでは計算ノード数に比例してデータ転送量の総和が増加するのに対し、提案手法では計算ノードが増加した場合においてもデータ転送量の総和は変わらないことを示している。

以上により、本論文の成果は計算機によるゲノム解析研究の発展に貢献するものと考えられる。よって博士（情報科学）の学位論文として価値あるものとして認める。