氏名（本籍）	井　上　悦　子（大阪府）
学位の種類	博　士（工学）
学位授与番号	甲 第１１号
学位授与日付	平成１９年3月２３日
専　　攻	システム工学専攻
学位論文題目	大規模アグリバイオデータからの知識発見を効率化するデータ解析支援システム
学位論文審査委員	（主査）教　授　中　川　　優 （副査）教　授　宗　森　　純 講　師　村　川　猛　彦

論文内容の要旨

本論文では，筆者が和歌山県地域結集型共同研究事業に携わり，プロジェクト内の3つのテーマに対して，大規模データからの知識発見作業を効率化するためのソフトウェアについて述べた．構築した2件のデータベースシステムについては，研究者の実験を支援するという立場から，各生物種の専門家から詳細なヒアリング調査を行い，個別の各研究においてどのようなことが望まれているかを把握した上で，研究を効率化し，かつ使用ストレスのないシステムの設計・構築を目指した．一方，大規模クラスタリング結果の可視化システムは網羅的解析全体を支援するものであり特定の実験を対象とするものではないが，これも研究者とのそれまでの活動を通じて得られた知見などを基に，現場での分析作業を効率化する使いやすいツールとしての実装を目指した．

まず，トランスポゾンを利用した遺伝子機能解析のための実験を支援するデータベースシステムを構築した．この実験は，トランスポゾンと呼ばれるゲノム上を自由に動くことのできる遺伝子を用いて突然変異を発生させ，様々な表現型（観測可能な性質）の変化との変異遺伝子との関係を調べることで，遺伝子機能を推定する実験手法である．この実験を効率化するために，DNA シーケンサによる遺伝子型の分析結果から検出される変異情報と圃場で計測される表現型情報をすべてデータベース化した．その上で，従来手法では各サンプルの遺伝子型をそれぞれピークのある波形データとして印刷して目視により比較していたものを，システム画面上で重ね合わせて比較できるようにすることで，変異箇所の特定の手間を大幅に削減し，実作業量にして十数倍の効率化を達成した．さらに，比較する複数サンプルの変異箇所を Web ブラウザ上で登録し一覧表示できる機能や，変異と形質の関連性を検定により評価する機能を備えることで，トランスポゾンディスプレイ解析の一連の実験全体を扱うことのできる統合的なデータベースシステムとして完成することができた．評価としては，変異データの重ね合わせ精度，実使用における使用感，時間短縮の効果についての評価を行い，本システムの有用性を示した．本システムを用いることで，現場での分析作業を従来の十分の一以下にまで大幅に短縮でき，継続してトランスポゾンディスプレイによる生物学的成果が出るなど，実質的な成果を挙げることに成功した．

次に，全国各地の農業試験場で実施されている種々の農業試験のデータを共有・管理し，データ傾向を効率的に把握するためのデータベースシステムを構築した．全国で行われている農業試験は，農作物の品種間のストレス耐性比較や肥料による生育度合いの比較など多様な目的で行われるが，取り扱う生物種や測定項目も多岐にわたるため，従来は担当の研究者が Excel などの表計算ソフトウェアを用いることで分析作業が行われてきた．この手法ではデータを担当研究者だけが管理するため研究グループ全体で共有できないことに加えて，データのおおよその傾向を把握するためにも手作業でのデータ整形・グラフ作成を繰り返し行う必要があり，手間がかかるだけでなく，複合目的の複雑な試験を行うことや試験データから予想外の現象に気づくことが難しいという課題があった．これらの課題を解決するために，筆者らはまず農業試験の典型的な分析パターンを検討し，これを基に試験データのモデル化を行った．また，多様な試験データをこのモデルに適合する形式でデータベース化することで，Web ブラウザ上で比較する条件や項目を選択する程度の単純な操作で，試験データの傾向把握に必要な典型的な分析を行うことができるようになった．なお，本システムは，現場の研究者にストレスなく使用できるように，従来どおりの Excel などの表計算ソフトを用いた簡単なデータ登録をはじめとした，細部にまで使いやすさに配慮したインタフェース設計を行っている．本システムの評価として，ウメとカキを用いた試験2例について現場の研究者の分析作業に試用してもらった結果，この分野では比較的複雑と思わ

れる試験内容であったにもかかわらず，問題なく適用可能であった．また，ヒアリング調査によりシステムの使用感を訊ねたところ，目的のために必要十分であり使いやすいとの良好な評価を得ることができた．さらに，和歌山県農林水産総合センターの研究報告書１年分の報告例(117 件)を用いてデータモデルおよびシステムへの適用可能性を調査し，大部分の試験に対して適用可能であることを確認した．これにより，データモデルの妥当性とシステムの有用性を示すことができた．システムの評価はまだ限定的であり今後広く多様な利用実験が求められるものの，適用事例に関しては有効な作業の効率化が見られ，現場での分析作業の省力化を実現できた．

最後に，バイオ分野の網羅的解析を支援するために，実験データに対する種々のクラスタリング結果をグラフにより表現し，表示範囲やデータ粒度を自由に変化させながら閲覧することができるインタラクティブな可視化方式を考案し，これを実現するソフトウェアを試作した．近年バイオ分野では，遺伝子やたんぱく質の発現量を網羅的に解析する例が増えており，数万～数十万というサイズの膨大な実験データをクラスタリングにより解析を行う重要性が増している．しかし，一般的な解析ソフトウェアでは，樹状図（デンドログラム）を用いたクラスタリング結果の可視化方法しか実現されておらず，データ同士の関係の詳細までを自由に閲覧することができない．これに対し，本手法では樹状図で用いられる木構造よりも密なネットワーク構造を用いて，表示画面中のノード間の関係を自由な範囲・粒度で閲覧できるため，従来手法に比べて格段に情報量が多く，バイオ研究者にとってより有用なデータ閲覧ツールになると考えられる．このような閲覧方式を実現するためには，指定された粒度や表示範囲に応じてノードを選択表示し，粒度を上げた場合にはノードの分割，粒度を下げた場合にはノードの結合をリアルタイムに画面上のグラフに反映してユーザの操作を追従できるインタラクティブな表示アルゴリズムを実現する必要がある．筆者は，数万程度の大規模データに対してもこのような連続的なグラフの変化を追従し，画面あたりのノード数とリンク数を一定の範囲内に制御できるデータ構造およびアルゴリズムを提案した．また，画面あたりの最大表示ノード数をＮとした時に，１回の粒度の変化に対して$O(N^2)$の計算量で追従可能であることを示し，クラスタリング結果のデータサイズに依存しない計算量での連続的な可視化処理を実現した．さらに，本アルゴリズムを実装したビューアソフトウェアを試作し，実際にデータ数１万のクラスタリング結果データの可視化に適用し，ストレスのない連続的な閲覧操作が可能であることを確認した．本研究の実用化はまだまだこれからであるが，膨大な発現量データの高速かつ効率的な新しい概観手法の提案を行い，その実現可能性を証明することができた．

これらの 3 つの研究活動は，バイオインフォマティクスの分野の中でも主に現場の研究者の研究活動を支援し効率化するものという位置づけで行ってきた．現場で様々な研究者や技術者と交流する中で，まだまだ生物の研究活動の中で現場レベルの作業を効率化するために情報技術を適用する余地が残っていることを実感した．近年は大規模データへの網羅的な分析が注目され，バイオインフォマティクス分野の発展が目覚ましいが，バイオインフォマティクスは主流の研究である情報技術による新たな解析手法の開拓だけでなく，取り扱う情報量が飛躍的に向上しつつある現場の作業を直接的に支援する方向にも研究の手を広げるべきであろう．本論文のテーマのうち，はじめの 2 つのシステムの設計・構築はそのような直接的な研究支援の典型例であり，このようなシステムへの試みが現場の研究効率を，1，2 割程度の向上にとどまらず，研究計画が変わりうるレベルにまで向上させ得ることを示している．また，最後のテーマはユーザとして情報系の分析担当者ではなく現場の生物学研究者を想定し，大規模データの可視化作業を生物学者が詳細に行えることによるデータ把握作業の効率化を狙ったものである．これらの 3 テーマの研究活動より，現場の研究者を支援することの有用性を示せたのではないかと思う．今後，本論文のような支援研究が広まり，現場の情報環境の改善による研究の効率化が進むことを心より願って止まない．

論文審査結果の要旨

論文の記述において、一部に誤記（括弧の抜け）および用語の説明不足が見られたが、容易に修正可能なこと、また、研究対象の技術間の繋がりを明確にする記述を追加することで論文のまとまりが良くなる意見もあり、若干修正することとした。

最終試験結果の要旨

約４０分間のプレゼンを堂々とこなし、また、審査員からの質問に対しても、的確に答えることができた。また専門分野での今後の活躍が期待できると判断する。