博士学位論文審査要旨

２０１０年2月１７日

論文題目：　大規模計算環境におけるタンパク質立体構造予測のための分布推定アルゴリズムの研究

学位申請者:　中尾　昌広

審査委員：

主　査：　工学研究科　教授　三木　光範

副　査：　生命医科学研究科　教授　廣安　知之

副　査：　工学研究科　教授　渡部　広一

要　　旨：

本論文では、タンパク質の立体構造を短時間かつ精度高く予測するために、「膨大な演算量を処理可能な計算環境に関する研究」および「性能の高い最適化アルゴリズムの開発に関する研究」を行った。

第３章では、膨大な演算量を処理するための計算環境として、複数の汎用 PC を LAN で接続した PC クラスタに着目し、PC クラスタを短時間にセットアップ可能なツールの開発を行った。さらに、PC クラスタの演算性能を高めるため、PC クラスタ内インターコネクトの高速化に関する研究を行った。VLAN 技術を用いて様々なネットワークトポロジを構築し、性能比較を行った結果、約 1100%の演算性能の向上を達成した。また、他の計算環境として、複数の PC クラスタを広域ネットワークで接続したグリッド環境にも着目した。グリッド環境を有効に利用するためには、アプリケーションがシステム情報を必要に応じて取得し、その情報をもとにアプリケーションの動作を柔軟に変更する仕組みが必要である。そこで、アプリケーションからグリッド環境の計算資源のシステム情報を取得可能にするミドルウェアの開発を行い、グリッド環境上で極めて低い負荷で動作することを示した。第４章ではタンパク質の立体構造を精度高く予測するために、最適化アルゴリズムの 1 つである分布推定アルゴリズムに着目し、その改良を行った。タンパク質の立体構造は局所的に安定する構造が多数存在するため、解の収束の速度を調節する新しい分布推定アルゴリズムを提案した。まず提案したアルゴリズムの性質を調査するため、数学的テスト関数を用いて数値実験を行った。その結果、提案アルゴリズムは既存の最適化アルゴリズムと比較して性能が高いこと、またタン

パク質の立体構造予測に適していることを示した。第５章では、第４章で提案したアルゴリズムをタンパク質の立体構造予測問題に適用した。その結果、既存アルゴリズムと比較して精度の高いタンパク質立体構造を得られることを示した。また、提案アルゴリズムの計算時間を短縮するために、PC クラスタ上で動作する並列モデルを開発した。分布推定アルゴリズムの解の評価計算を複数の計算ノードで分散処理することで、PC クラスタを用いない場合と比較して計算時間を約 1/35 にまで削減できることを示した。

本論文の成果により、PC クラスタおよびグリッド環境を簡易に構築、利用できるようになった。さらに、提案アルゴリズムを用いることで、精度が高いタンパク質立体構造の予測を行うことが可能になった。この成果は、タンパク質の構造解析分野において多大な恩恵をもたらすことが期待できる。

よって、本論文は、博士（工学）（同志社大学）の学位論文として十分な価値を有するものと認められる。

総合試験結果の要旨

２０１０年2月１７日

論 文 題 目：　大規模計算環境におけるタンパク質立体構造予測のための分布推定アルゴリズムの研究

学位申請者:　中尾　昌広

審 査 委 員：

主　査：　工学研究科　教授　三木　光範

副　査：　生命医科学研究科　教授　廣安　知之

副　査：　工学研究科　教授　渡部 広一

要　　旨：

本論文提出者は 2007 年 4 月より本学大学院工学研究科博士課程後期課程に在学している。本論文の主たる内容は、2007 年度および 2008 年度に、日本計算工学会論文誌に 1 報ずつ、情報処理学会論文誌コンピューティングシステムに 1 報、掲載されている。また、査読付き国際会議の論文として、1 報掲載されている。各年度において優れた研究成果を挙げ、英語の語学試験に合格し、ハングル語についても十分な能力を有すると認定されている。本年 1 月 9 日に開催された博士論文公聴会においては、十分な学力、将来性、研究の深さを確認した。

よって、総合試験の結果は合格であると判定した。

博士学位論文要旨

論文題目：大規模計算環境におけるタンパク質立体構造予測のための分布推定アルゴリズムの研究

氏　　名：中尾　昌広

要　　旨：

タンパク質は20種類のアミノ酸が鎖状に連結した物質であり，生命の重要な構成物質の1つである．タンパク質の持つ機能的性質は立体構造と密接な関係にあるため，アミノ酸配列に対応するタンパク質立体構造を解明することにより，新薬創製促進や病理原因解明などの効果が期待できる．本論文の主題は，最適化手法を用いたコンピュータシミュレーションによるタンパク質の立体構造予測である．最適化とは，定式化した対象問題の最小値（又は最大値）を与える変数を決定することを指し，工学のみならず経済学などの様々な分野で研究が行われている．タンパク質は特定の形に折り畳まれたエネルギーが最も低い状態で存在するため，タンパク質のエネルギー関数を定式化し，最小化を行うことで立体構造を予測することが可能である．しかしながら，最適化アルゴリズムを用いてタンパク質の立体構造予測を行うには膨大な演算量が必要であるため，現実的な時間で計算を行うための大規模計算環境の構築が必要である．さらに，タンパク質のエネルギーを最小化する性能の高い最適化アルゴリズムの開発も必要である．このような背景から，本論文は「大規模計算環境を短時間かつ簡易に構築，管理を行うためのソフトウェアの開発」および「大規模計算環境で動作する性能の高い最適化アルゴリズムの開発」を行うことで，タンパク質の精度の高い立体構造予測を試みる．

まず，「大規模計算環境を短時間かつ簡易に構築，管理を行うためのソフトウェアの開発」について述べる．有力な大規模計算環境としては，複数の汎用PCをLANで接続したPCクラスタ，また複数のPCクラスタを広域ネットワークで接続したグリッド環境が考えられる．PCクラスタを構築するソフトウェアは数多く存在するが，既存ソフトウェアの問題点として，操作には専門的知識が必要である点，PCにハードディスク（HDD）を搭載せずに動作するディスクレスノードとHDDを搭載して動作するディスクフルノードの混合環境の構築が行えないという点が存在する．特に混合環境の構築は導入コスト削減，故障率低下および省電力の面から実現が強く望まれている．そこで，PCクラスタの専門知識を持たないユーザでも扱うことができ，かつ混合環境を構築することができる自己組織型PCクラスタ構築ツール（Dynamic Cluster Auto Setup Tool：DCAST）の開発を行った．DCASTはネットワークブートの仕組みを利用した

インストーラであり，ユーザはPCクラスタを構成する1台のPCにOSとDCASTをインストールするのみで，すべてのPCに対してPCクラスタの設定を行うことができる．また，DCASTはPCクラスタを構築する際に各ノードのHDDの有無を判別し，HDDを搭載しないノードはディスクレスノード，HDDを搭載するノードはディスクフルノードとして構築を行う．さらに，DCASTではディスクレスノードとディスクフルノードとの対応関係を均等に結ぶことで，ネットワークの負荷分散を自動で行う機能を開発した．本論文ではDCASTの実証実験を行い，約200ノードで構成されたPCクラスタを1時間程度で構築可能であることを示した．また，計算機管理能力が異なる被験者10名に対し，PCクラスタの構築時間とPCクラスタ時におけるユーザの負担について調べた．その結果，DCASTは既存ソフトウェアと比較して構築時間とユーザの負担が有意に少ないことを示した．

次にグリッド環境の利用を考えた場合，ユーザは利用できる計算資源のシステム情報（ノードの性能，負荷率など）をアプリケーション実行前に入手する必要がある．しかし，グリッド環境の計算資源は頻繁に変化し，メンテナンスや故障による削除も行われる．そのため，グリッド環境を有効に利用するためには，アプリケーションがシステム情報を必要に応じて取得し，その情報をもとにアプリケーションの動作を柔軟に変更する仕組みが必要であると考えた．しかし，そのような仕組みを実現するソフトウェアは存在しない．そこで，アプリケーションからグリッド環境の計算資源のシステム情報を取得可能にするミドルウェアである分散ネットワークアプリケーションシステム（Distributed Network Application System：DNAS）の開発を行った．DNASはノードのシステム情報を定期的に取得し，その情報をユーザが利用できる形（API）で提供する．ユーザはDNASが提供するAPIを用いたソフトウェアを作成することで，計算資源の動的変化を考慮したアプリケーションの開発が可能になる．例えば，数値計算アプリケーションを実行する場合，ノードの変化にともない動的にパラメータを変化させることで，効率的に計算を行うことが可能になる．また，グリッド環境には多くの計算資源が存在するため，1台のノードにすべてのシステム情報を集約することは，可用性の点からも好ましくないと考えた．そこでDNASにP2Pの通信形態を実装し，各ノードは別のノードと対応関係を結ぶことでシステム情報の通信を相互に行い，その情報を保存する機能を実装した．本論文ではDNASの実証実験として，3つのPCクラスタと単体のPC（計230ノード）を用いた大規模グリッド環境においてDNASの動作を確認した．その結果，DNASは極めて低い負荷で動作し，かつ障害に強い情報通信網が確立することを示した．

次に，「大規模計算環境で動作する性能の高い最適化アルゴリズムの開発」について述べる．最適化分野においてタンパク質の立体構造予測は非常に困難な問題の1つであ

るとされている．その理由は，タンパク質のエネルギー局面（ランドスケープ）はいたる所に局所安定構造があり，かつタンパク質のエネルギー関数が持つ変数はそれぞれ依存しているなどの特性があるからである．そこで，近年注目されている最適化アルゴリズムの 1 つである分布推定アルゴリズム（Estimation of Distribution Algorithm : EDA）に着目し，その改良を行った．EDA は複数の有望な候補解の統計情報から確率モデルを構築し，その確率モデルから新しい解を生成するというアルゴリズムである．EDA は用いる確率モデルによっては変数間に依存関係を持つ問題にも対応でき，また極めて汎用的であるという特徴を持つ．本論文では，EDA の解が局所安定構造に収束することを防ぐため，解の収束速度が異なる複数の確率モデルを用いた新しい EDA である複数の確率モデルを用いた実数値分布推定アルゴリズム（Real-coded EDA Using Multiple Probabilistic Models : RMM）を開発した．RMM の性能を調べるため，数学的ベンチマークテスト問題を用いて性能を測定した．その結果，既存の EDA と比較して RMM の性能が高いことを示した．さらに，変数間に依存関係を持つような問題にも対応させるため，RMM の解に対して独立成分分析を行うことでランドスケープに沿った解を生成する機能を開発した．この機能を用いることで，依存関係を持つ問題に対して，より RMM の性能が高くなることを示した．

次に，複数の確率モデルを用いた実数値分布推定アルゴリズム RMM をタンパク質の立体構造予測問題に適用させ，その解探索メカニズムを詳細に検討した．その結果，新しい解を生成した際にタンパク質の部分構造同士が頻繁に衝突するという問題点が明らかになった．そこで，タンパク質の構造を少しずつ変化させ，さらに新しい解を生成した際に局所探索を行うことで，部分構造同士の衝突を回避する工夫を行った．また，一般的な EDA では確率モデル構築に用いた解は新しい解と置換されるため，再利用されることはない．そのような世代交代モデルでは，良い構造を持つ解も置換されるため，タンパク質の立体構造予測問題においては好ましくないと考えた．そこで新しい解を生成した際に，確率モデル構築に用いた解と新しい解との比較を行い，良い評価値を持つ解を次世代に残すという工夫を行った．これらの工夫を行った結果，RMM は性能の高い最適化アルゴリズムである PSA/GAc と比較してタンパク質の精度の高い立体構造を得られることを示した．

最後に，　タンパク質の立体構造予測を行う計算時間を削減するために，複数の確率モデルを用いた実数値分布推定アルゴリズム RMM の並列化アルゴリズムを開発した．予備実験から，RMM はタンパク質のエネルギー計算と局所探索に多くの処理時間を要していることがわかった．そこで，タンパク質のエネルギー計算と局所探索を異なる CPU コアで処理する並列アルゴリズムを開発した．計算に用いる CPU コア数と計算時間の関係を調べた結果，並列化 RMM は CPU コア数に対してほぼスケールに計算時

間を短縮できることを示した．具体的には，1CPU コア時と比較して計算時間を約 1/80 にまで削減できることを示した．

本論文の成果により，PC クラスタおよびグリッド環境を簡易に構築，利用できるようになった．さらに，提案する最適化アルゴリズムを用いることで，タンパク質の精度が高い立体構造の予測を行うことが可能になった．この成果は，タンパク質の構造解析分野において多大な恩恵をもたらすことが期待できる．