

大規模タンパク質全電子計算を目指した 地球シミュレータ版 ProteinDF の開発

吉廣保[†]、稲葉亨^{‡*}、上野哲哉[†]、佐藤文俊[†]、柏木浩^{†‡}

東京大学生産技術研究所[†] アドバンスソフト株式会社[‡] 株式会社日立製作所^{*}

1. はじめに

生体内で重要な働きをするタンパク質は多くの原子からなる生体高分子である。特に大型のタンパク質においては、基本骨格を作る 20 種類のアミノ酸以外にも、ヘムやクロロフィルに代表されるような色素分子が複雑に入り込んで、機能性に富んだタンパク質固有の反応を発現している。タンパク質の機能を理論的に電子レベルで解明するためには、タンパク質全体を分子として扱うことができ、かつ多くの色素に含まれる金属もアミノ酸残基と同程度の精度で求めることができる解析手法が望ましい。

そのような観点から、発表者らはこれまで独自にガウス型基底関数を用いた密度汎関数法に基づく大規模分子軌道法プログラム ProteinDF を作成・発展させてきた[1]。密度汎関数法は Kohn-Sham 方程式に基づく分子軌道法の一つで、最大の利点は *ab initio* HF 法と同程度の計算量で、電子相関効果を取り込んだ計算ができることである。そのため、大型のタンパク質の計算に威力を発揮するであろう。また、開発したプログラムは、カノニカル分子軌道計算においても、計算時間が軌道数の約 2.4 乗で計算できるという良好な結果を得ている[2]。

2. オブジェクト指向技術による並列化と WS クラスタによる並列分散処理

ProteinDF プログラムは大規模な行列等のデータを安全に取り扱い、かつプログラムを多人数で共同開発することを念頭に置き、オブジェクト指向言語 C++ でコーディングされている[1]。

発表者らは、オブジェクト指向の考え方を並列化にも導入して、プログラム全体をアルゴリズム、計算、通信といった各機能が互いに独立した構造となるよう設計したうえで、シナリオ（アルゴリズム）部、（並列）計算部、仮想マシン部、通信部に完全に階層化し、並列化した（図 1）。これらの階層は、機能的に全て独立しているため、計算機アーキテクチャや通信用ライブラリの多様性に容易に対応できる。また、各計算ルーチンが十分に独立しているため、他の計算ルーチンのデータ構造などを考えることなく個別に修正・改良を加えることができる。

この並列版 ProteinDF を用いて、HP ALPHA 21264 667MHz ワークステーション（WS）と同 21164A 533MHz WS 8 台、500MHz WS 6 台の計 15 台を 100Base-TX の Ethernet で接続した WS クラスタ並列システムを使用し、馬心臓シトクロム *c*（104 残基+ヘム *c*、1738 原子、9600 軌道、17578 補助基底関数）の全電子計算に成功した実績がある。しかし、このシステムでは繰り返し計算の 1 回にかかる実時間は 77362 秒（21.5 時間）で、59 回の繰り返し計算を要したため[2]、それほど手軽に実行できるわけではない。また、さらに大きなタンパク質の計算を行うには、この程度の計算機資源では不十分である。

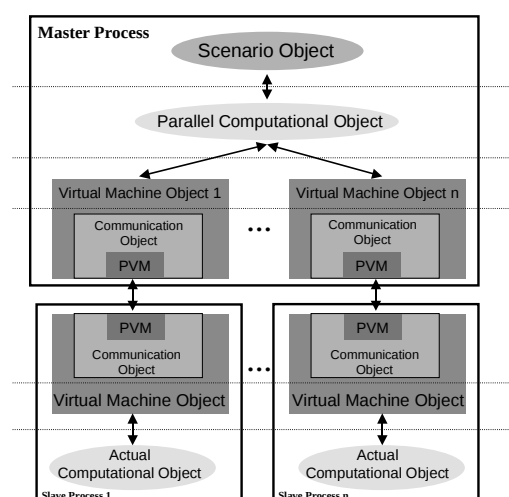


図 1: ProteinDF の並列化構造

3. 地球シミュレータへの移植とチューニング

そこで、本研究では地球シミュレータのような超並列大型計算機において ProteinDF が効率良く実行できるプログラムを開発している。本研究により、100 残基から数百残基のタンパク質

の全電子計算が、現実的な時間で実行できる基盤が整備できるものと考えられる。

地球シミュレータは、640 台の計算ノードを、640×640 の単段クロスバネットワークで結合させた分散メモリ型並列計算機である。各計算ノードは、ピーク性能 8Gflops のベクトル型計算プロセッサ 8 台が主記憶装置 16GB を共有する共有メモリ型並列計算機となっている。全体では計算プロセッサが 5120 台でピーク性能は 40Tflops、主記憶容量は 10TB となる計算機である。

このような計算機資源で ProteinDF が効率よく実行できるかを図る目安として、地球シミュレータに比べ、CPU の性能に対して遥かに低速なネットワーク性能を持つ WS クラスタ並列システムにおいて、どの程度の効率で ProteinDF が並列化されているかを調べた。計算には 31 残基ペプチド鎖断片 (2867 軌道、5273 補助基底関数) の計算を使用した。HP Alpha 21164A 533MHz WS 8 台を 100Base-TX の Ethernet で接続した 1, 4, 8 台のシステムを使用して、最初の SCF 計算にかかる実時間を測定し、1 台で計算した時間を基に使用台数による並列化効率を求めた。その結果、並列化効率は 4 台で 91.0%、8 台で 78.1%であった[3]。解析の結果、並列化効率を下げている最も大きな要因は行列データの転送時間であり、計算処理は 95%以上の効率で均等化されていることが明らかとなった。行列の転送時間よりも計算処理時間のほうが計算規模に対する依存性が大きいと、さらに大きな系ではより高い効率が見込まれる。これらの結果により、地球シミュレータにおいて高速なシミュレーションが達成できるものと考えられる。

地球シミュレータでの並列処理は、OpenMP や Message Passing Interface (MPI)を使用して並列化を行う。しかし、OpenMP も MPI も SPMD 方式の並列処理に優れた処理環境やライブラリである。並列版 ProteinDF は巨大計算のためにメモリをより柔軟に利用できるよう、マスタ・スレーブ方式を採用している。マスタ・スレーブ方式でプログラムをコントロールするためには動的プロセス制御を行える通信ライブラリが適しており、これまで Parallel Virtual Machine (PVM)を採用してきた。地球シミュレータにおいても、数万次元の行列を扱うには、計算ノード当たり 16GB のメモリを効率よく使用する必要があるため、本研究では、ProteinDF の地球シミュレータへ移植にも、あえてマスタ・スレーブ方式を採用し、動的プロセス生成のできる MPI-2 通信ライブラリを用いた。

通信クラスの MPI-2 通信ライブラリへの対応の際、通信クラス以外の部分においても一部 MPI 通信でのメッセージサイズを明示的に指示するための変更が必要となったが、図 1 のプログラム構造のお陰で、通信クラス以外にはほとんど手を加えることなく、PVM 版通信クラスと MPI-2 版通信クラスのどちらでも動作するプログラムを作成することが出来た。また、本研究における、現行の地球シミュレータの利用形態は、計算ノード内・計算ノード間を特に区別せずすべての計算プロセッサを対等に扱った、first-come/first-serv タイプのスレーブ計算処理スケジューリングによる並列処理法を採用している。

地球シミュレータはベクトル型計算プロセッサの並列マシンである。ProteinDF は分子積分や行列演算などが律速ルーチンである。各ルーチンは効率よく並列化されたので、続いてスレーブプロセッサで実行される実計算ルーチンのベクトル化を行っている。現在のところ、分子積分ルーチンはベクトル化率約 92%で、行列計算は専用ライブラリをスレーブプロセスで用いることで性能が得られている。

発表では、さらにチューニングを施した上、シトクロム *c* の全電子計算について WS クラスタと地球シミュレータでの計算結果をもとに、並列化効率などについて報告する予定である。

謝辞

本研究は文部科学省 IT プログラム「戦略的基盤ソフトウェアの開発」において実施された。さらに、地球シミュレータセンターに感謝します。

参考文献

- [1] F. Sato, Y. Shigemitsu, I. Okazaki, S. Yahiro, M. Fukue, S. Kozuru, H. Kashiwagi, *Int. J. Quantum Chem.* **63**, 245 (1997).
- [2] F. Sato, T. Yoshihiro, M. Era, H. Kashiwagi, *Chem. Phys. Lett.* **341**, 645 (2001).
- [3] T. Yoshihiro, F. Sato, H. Kashiwagi, *Chem. Phys. Lett.* **346**, 313 (2001).