平成 22 年度 RICC 利用報告書

課題名（タイトル）：

タンパク質・核酸など生体高分子の分子シミュレーション

利用者氏名：　○木寺　詔紀　　寺田　透　　森次　圭　　松永　康佑

所属　　：　社会知創成事業　次世代計算科学研究開発プログラム

次世代生命体統合シミュレーション研究推進グループ　分子スケール研究開発チーム

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

本研究チームでは、次世代生命体統合シミュレーション研究開発プログラムの分子スケール研究の一環として、生体分子（タンパク質等）のシミュレーション法とそのソフトウェアの開発研究、特に、全原子シミュレーション法と疎視化モデルとの新規の連成の方法論の開発を行っている。この研究の目的は以下の 2 点である：

1）次世代スーパーコンピュータの全計算機資源を用いて高効率で計算することができる

2）それによってこれまでの分子シミュレーションの方法ではできなかったレベルの計算をすることができる

生命活動をタンパク質や核酸などの生体分子のレベルからシミュレーションによって解こうという分野における問題は、その巨大な系の大きさと生命現象の時間スケールの大きさである。その大きさのために、全原子シミュレーション法には巨大な計算機資源を用いても多くの場合、生命現象の解明が可能な系の大きさと計算時間の長さを実現するシミュレーションは不可能である。そこで不可避的に疎視化モデルの利用が求められるが、そこには精度の制約が生まれる。従って、その両者の利点を併せ持つ連成計算（全原子シミュレーション法の精度と疎視化モデルの効率）が必要となる。また、数十万コアという並列計算を実現するためには、不可避的に弱連成のアルゴリズムであることが要請される。これらを可能とする新規アルゴリズムとして、経路探索法（On-the-fly ストリング法）と MultiScale Enhanced Sampling (MSES 法)の研究開発を行った。On-the-fly ストリング法では、タンパク質の構造変化パス上に配置し、それらマルチコピーを弱連成で並列に計算することで、最適経路をサンプルする。MSES 法では、全原子モデルの運動空間を連成した粗視化モデルでドライブすることにより全原子モデル巨大系の高速なサンプリングを実現する。

これらふたつの方法は、系が巨大タンパク質系であればそれだけで多くのコアによる計算を必要とし、さらにマルチコピーを必要とするところから、それらの掛け算の巨大な計算機資源を用いる必要がある。また、コピー間は弱連成であるところから次世代スーパーコンピュータの全計算機資源を用いて高効率で計算し得る可能性を持つ。これらの方法によってこれまでに全原子モデルシミュレーションでは不可能であった、長時間現象を再現することが可能となった。また、疎視化モデルでは不可能であった、全原子モデルの精度を与えることができる。

２．　具体的な利用内容、計算方法

On-the-fly ストリング法と MSES 法を、寺田が開発を統括しているマルチコピー・マルチスケール分子動力学シミュレーションソフトウェアに実装した。マルチコピーシミュレーションでは、異なるパラメータを与えた数十の系のコピー（レプリカ）を発生させ、それらの間の相互作用を考慮しながら並行してシミュレーションを行う。本年度の特別利用では、On-the-fly ストリング法と MSES 法について、flat MPI とハイブリッド並列でのスケーラビリティテストを RICC の上限である 8192 コアまでコア数を増やしながら実行した。

３．　結果

（１）マルチコピー・マルチスケール分子動力学シミュレーションソフトウェアの開発

次世代計算科学研究開発プログラム分子スケール研究開発チームでは、次世代スーパーコンピュータの計算能力を最大限に活用して、生体高分子の立体構造形成・機能発現メカニズムの解明を高精度かつ効率的に行うことのできる、新規アルゴリズムの開発を行って

いる。寺田、森次、松永は、これら新規アルゴリズムを実装するためのプラットフォームとなる、マルチコピー・マルチスケール分子動力学シミュレーションソフトウェアの開発を行っている。本ソフトウェアは MPI を用いて並列化されていたが、今年度は、並列計算効率の更なる向上を目指して、OpenMP と組み合わせたハイブリッド並列化を行った。本ソフトウェアでは、各コピー内の相互作用の計算を分割することにより、並列化を行っているため、相互作用の計算が完了するごとに、すべてのプロセスが計算結果を共有する必要がある。ハイブリッド並列化により、MPI のプロセス数を 1 プロセス／コアから、1 プロセス／ノードに減らすことができる（RICC の場合プロセス数は 1/8 になる）ため、計算結果の共有にかかる時間を大幅に削減することができた。RICC では、スレッド生成・消滅にかかるオーバーヘッドが大きいため、計算速度全体の向上には至っていないが、次世代スーパーコンピュータではオーバーヘッドが小さくなることが期待されており、通信時間の大幅な減少は、計算速度や並列計算性能の向上につながると期待される。

（2）ストリング法計算の高度化と adenylate kinase への応用

タンパク質機能にとって重要な立体構造変化はミリ秒～秒で起こる遅い過程であり、Brute force シミュレーションでこれを再現するのは困難であると同時に、仮に再現し得たとしても、構造変化の軌道から遷移状態周りがサンプリングされる割合は稀であるので、そこから構造変化の機構に関する知見を得ることは難しい。そこで我々は、単一のシミュレーションの代わりに、多数のコピー系をタンパク質の構造変化パス上に配置し、それらマルチコピー系を並列に計算することで粗視化空間における最適パスをサンプルする手法（ストリング法、L. Maragliano and E. Vanden-Eijnden, *Chem. Phys. Lett.* **446**, 182 (2007)）の実装・開発を行った。本年度特別利用では、開発したプログラムの大規模並列でのベンチマーク計測・プログラムの高度化とプロダクションランを行った。

以下簡単にプログラムの説明を行う。まず、プログラムは将来的にユーザがマルチコピー系シミュレーションを容易に実現でき、自由にオリジナルのシミュレータを構築できるように、オブジェクト指向に基づいて設計しており、C++でコードされている。マルチコピー系の計算のために、MPI communicator を MPI_Comm_split 関数で分割することで、コピー内・コピー間通信のための communicator をそれぞれ準備している。コピー内では、原子分割で MPI 並列化しており、長距離相互作用項の計算のために smooth particle mesh Ewald 法(SPME)を実装している。年度後半には、更にペアリストのループ部分などを分割・OpenMP 化し、ハイブリッド並列実行もできるようにした。コピー間では、他のコピー系とのバネ的な相互作用を評価するために、各コピー系のマスタープロセス同士で MPI 通信を行う。具体的には、各コピー系がそれぞれ保持するタンパク質の粗視化変数が、粗視化空間の構造変化パス上で等間隔になるような操作を行う。等間隔にするためには、パスの全長情報が必要なので、MPI 集団通信を行っていたが、これだと通信遅延が大きいことが判明した（表 1）。そこで年度途中に、隣接通信の繰り返し処理で近似するアルゴリズムを開発した（表 2）。

以下、ベンチマーク計測結果を報告する。計算対象は、adenylate kinase (3,343 原子)と水分子(TIP3P モデル)・イオンから成る計 62,475 原子系であり、コピー数を 256 とした。したがって、1,024 コアでは 4 コア/1 コピー、2,048 コアで 8 コア/1 コピー、4,096 コアで 16 コア/1 コピー、8,192 コアで 32 コア/1 コピー、となる。表 1 に、flat MPI 版のベンチマーク結果を示す。Setup 項は座標やトポロジー情報をファイルから読み込んで MPI_Bcast する時間、Ewald 項は SPME の計算時間、String 項はコピー間相互作用の計算時間、Force communication 項はコピー内の各 MPI プロセスにおいて計算された force を MPI_Allreduce 通信する時間を計測している。表 1 を見ると、計算量が大きい Ewald 項の時間が最も大きく、次いで、Force communication と String 項が大きいのが分かる。特に後者 2 つは、コア数の増大とともに、時間が増えており、8,192 コアでスケーリングが悪くなる原因となっている。そこで、前記のとおり String 項計算のアルゴリズムを集団通信から隣接通信の繰り返しに置き換えたアルゴリズムを開発し再度ベンチマーク計測を行った(表 2)。表 2 では、String 項がわずかにではあるが改善されていることが分かる。次に、OpenMP 化によるハイブリッド並列実行による Force communication 項の変化を調べた（表 3）。表 3 では、ハイブリッド並列化により、Force communication

項が改善されているのが分かる。前記のとおり、Force communication 項では、各 MPI プロセスで計算した force を MPI_Allreduce 通信を使ってまとめるということをしている。 それをハイブリッド並列化によって、より少ない MPI プロセスによる MPI_Allreduce 通信に置き換えたために、つまり通信量を削減したために改善されたと考えられる。ちなみに、ハイブリッド並列化版(表 3)では、スレッド生成のオーバーヘッドのために 1,024～4,096 コアで flat MPI 版(表 1)より total 時間が遅くなっているが、8,192 コアではより速くなっている。今後、次世代スーパーコンピュータ「京」などの更に大規模な並列環境では、ハイブリッド並列化版のほうが良い性能が期待できると考えられる。

Cores	1024	2048	4096	8192
Setup	1.29	1.23	1.20	0.98
Bond	1.03	0.73	0.41	0.22
Angle	0.19	0.14	0.10	0.06
Dihedral	1.42	1.02	0.59	0.38
Ewald	442.30	319.64	189.70	107.86
SHAKE	1.14	1.35	1.41	1.41
Restraint	0.05	0.04	0.04	0.02
String	22.55	25.10	29.34	33.71
Force communication	11.69	18.73	34.21	53.37
Total	515.25	393.16	273.08	208.38
Speedup	1.00	1.31	1.89	2.47

表 1 Flat MPI でのストリング法計算ベンチマーク結果(秒)

Cores	1024	2048	4096	8192
Setup	1.24	1.16	1.02	0.90
Bond	1.03	0.73	0.41	0.22
Angle	0.19	0.14	0.10	0.06
Dihedral	1.42	1.02	0.59	0.38
Ewald	443.00	320.28	189.84	107.92
SHAKE	1.16	1.36	1.41	1.41
Restraint	0.05	0.04	0.04	0.02
String	11.07	13.46	20.99	26.32
Force communication	11.69	18.72	34.50	53.39
Total	504.39	382.06	264.98	200.99
Speedup	1.00	1.32	1.90	2.51

表 2 Flat MPI、隣接通信版のストリング法計算ベンチマーク結果(秒)。隣接通信の繰り返し回数は 5 回とした。

Cores	1024	2048	4096	8192
Setup	1.43	1.49	1.25	1.12
Bond	1.25	0.97	0.58	0.30
Angle	0.25	0.19	0.13	0.08
Dihedral	1.58	1.26	0.72	0.44
Ewald	486.20	362.90	211.10	116.23
SHAKE	0.78	1.02	1.03	1.00
Restraint	0.06	0.07	0.07	0.06
String	13.35	35.75	36.50	30.65
Force communication	1.49	2.26	4.97	8.23
Total	550.33	455.33	293.48	181.67
Speedup	1.00	1.21	1.88	3.03

表 3 Hybrid 並列版のストリング法計算ベンチマーク結果(秒)。1,024 コアでは 4 スレッド×256 プロセス、他は全て 8 スレッドとした。

（3）MSES 法でのスケーラビリティテスト

森次は sortase（溶媒水を含めて約 3 万原子）をテスト系として MSES 法のスケーラビリティをテストした。ベンチマークとして、1,000 ステップの分子動力学シミュレーションを実行した。Strong scaling では 512 コピーの計算をコア数を変えながらで実行し、また、weak scaling では 1 コピーにつき 16 コアに固定してコア数を変えながら実行した。

その結果、strong scaling、weak scaling ともに RICC の全コア数である 8192 コアまでよくスケールすることを確かめた。また、ハイブリッド並列のテストを実行し flat MPI の結果と比較した結果、コピー内の通信である force communication の時間が大幅に減少できていることがわかった。

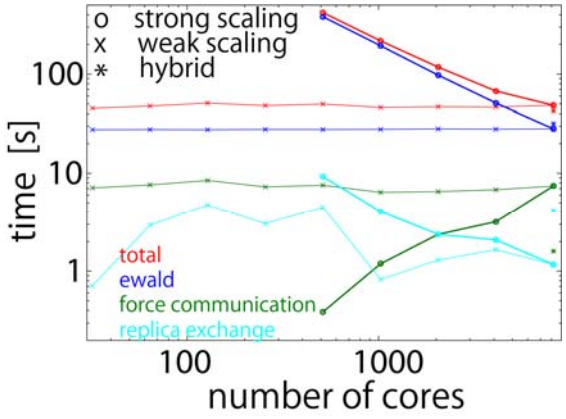

図 2: MSES 法のスケーラビリティ。Strong scaling は 512 コピーの計算をコア数を変えながらで実行。weak scaling は 1 コピーにつき 16 コアに固定してコア数を変えながら実行。

４． 今後の計画・展望（これまで利用した状況や、継続して利用する際に行う具体的な内容）

マルチコピー・マルチスケール分子動力学シミュレーションソフトウェアの開発をすすめるため、新規アルゴリズムの開発において追加される機能の追加、入出力などのクラスライブラリ高度化、また、超並列マシンである次世代スーパーコンピュータへの対応として OpenMP と組み合わせたハイブリッド並列化によるコードの高速化をさらに進めていく。

平成 22 年度　RICC 利用研究成果リスト

【論文、学会報告・雑誌などの論文発表】

1.　Kei Moritsugu, T. Terada and Akinori Kidera, “Scalable Free Energy Calculation of Proteins via Multiscale Essential Sampling”, *J. Chem. Phys*. 133, 224105 (2010).

【国際会議などの予稿集、proceeding】

該当なし

【国際会議、学会などでの口頭発表】

１．　森次 圭

「MSES シミュレーションによる Sortase A 不規則領域の構造空間探索」

次世代生命体統合 PJ　分子スケール研究会 2010 前期、　京都、2010 年 8 月

２．松永 康佑、寺田 透、森次 圭、古田 忠臣、木寺 詔紀

「マルチコピーシミュレーションによるタンパク質構造サンプリング」

2010 年度理研シンポジウム　ペタフロップス時代のセンターシステム、和光、2011 年 2 月

RICC カッティングエッジアワード受賞

３．木寺 詔紀

"Supercomputing of protein systems using innovative algorithms for multiscale simulations"

バイオスーパーコンピューティング研究会　講習会、大阪、2011 年 3 月

【その他】

該当なし