Activity Report on Bio Molecular Simulation Consortium in 2004

Consortium Representative
Toshikazu Takada Fundamental and Environmental Research Laboratories NEC Corporation

Author
Toshikazu Takada Fundamental and Environmental Research Laboratories NEC Corporation

1. Outline of Consortium

Bio Molecular Consortium was established in March, 2003. The main mandate of the consortium is to evaluate proposals from the members to Earth Simulator among themselves from professional point of view and then to recommend proper proposals with certification of the consortium. As mentioned below, through the evaluations, those proposals are brushed up to have a high quality of research as projects on Earth Simulator. The number of the members is now more than 40 and they are the top scientists in this field of Japan.

2. Goals of Consortium

As an important research area, bio technology is being recognized recently. Thus, our objectives are:

1) To understand the mechanism of biological systems at molecular levels
2) To apply the knowledge obtained from simulations to industrial activities such as drug and enzyme design

Basically, all the codes of bio molecular simulations are originally developed in order to achieve these objectives with high performance on Earth Simulator and related simulations are carried out to demonstrate their usefulness for the second objective.

Another is to recommend proposals to Earth Simulator, which are proper projects to be carried on it. To be distinct, these proposals are sent to all the consortium members for peer review. Then, after such selection, the chair will recommend them with certification of the consortium. Therefore, the research quality of the proposals which are recommended, are kept to be high from professional point of view.

3. Activities of Consortium

As mentioned before, the major activity of this consortium is to recommend bio-related material simulations to Earth Simulator. The other is to provide these researchers with a place for sharing the information of simulations related to biology, since these areas have been developing rather independently. For detailed understanding of the mechanism of biological systems, collaborations by these researchers of different areas such as genome informatics, protein folding, molecular dynamics and so on, are needed. In the consortium, this is achieved by circulating their proposals before applying to Earth Simulator.

4. Sub-Themes of Project

1) “All-electron Calculation on Very Large-Sized Proteins by Density Functional Method”, Fumitoshi Sato, Institute of Industrial Science, University of Tokyo
2) “Realistic simulations of the structural changes of proteins”, Minoru Saito, Faculty of Science and Technology, Hirosaki University
3) “Protein folding simulations from the first principles”, Yuko Okamoto, Department of Theoretical Studies, Institute for Molecular Science
4) “The Molecular Dynamics Simulations of Prion Protein: Investigation of the Transition from its Cellular Form to the Anomalous Form using Earth Simulator”, Yutaka Akiyama, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)
5) “Analysis of the function of a large-scale supra-biomolecule system by molecular dynamics simulation”, Hisashi Ishida, Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute
6) “Self-Organizing Map (SOM) for all genome and protein sequences and simulation of the genome evolution”, Toshimichi Ikemura, Hayama Center for Advanced Research, the Graduate University for Advanced Studies (Sokendai)

5. Cross Relationship of Sub-Themes

To understand mechanism of biological systems at molecular levels, it is needed to hybridize many of individual theories and computational technologies which have been developed as separated areas such as genome informatics, protein folding, molecular dynamics, electronic structure and so on. Therefore, Bio Molecular Simulation Consortium now conducts 6 sub-themes as the Earth Simulator projects shown below.

<Sub-Theme 1>

Here, a gaussian-based density functional simulation code named ProteinDF for study of electronic structures of proteins has been developed, which is able to handle a whole proteins as one molecule with more than 100 residues. To understand the mechanism of chemical reactions occurring in proteins, it is needed to study electronic structures of proteins. For this, the number of necessary basis functions is to be around 10,000. There is no molecular simulation code which handles such large molecular systems. ProteinDF is coded by C++. The purpose of this project is to carry out all-electron calculation on 1,000 residues complex protein on Earth Simulator.

<Sub-Theme 2>

In this sub-theme, a powerful molecular dynamics simulation algorithm called Replica-Exchange Molecular Dynamics (REMD) has been applied to the ab initio folding simulation of a rather small protein, protein G, in explicit water. The numbers of amino acids and surrounding water molecules are 56 and 17,187. The total number of replicas is 224 in this simulation using successfully 896 CPUs of Earth Simulator. A lot of secondary-structure formations were observed during the simulation, some of which are native-like. From now, more powerful simulation algorithm (MUCAREM) than REMD for folding simulation of protein G in explicit water will be tested in order to further enhance sampling.

<Sub-Theme 3>

The purpose is to computationally demonstrate and visualize the structural changes of hemoglobin using COSMOS90 which is able to efficiently simulate protein motions in water with all degrees of freedoms and long-range Coulomb interactions. The second stage of 2004 was data preparation of hemoglobin in a realistic environment and acceleration of COSMOS90. The performance speed of 0.023 sec/step for hemoglobin in water (120036 atoms) is successfully achieved, where all the sub-processes of COSMOS90 including Barnes-Hut tree code was vectorized and parallelized. This performance is found to be faster than any other software widely used in this research filed such as CHARMM, AMBER7 and NAMD2.4.

<Sub-Theme 4>

A main goal of prion protein research is detection of the process that causes the conformational change from normal cellular form (PrP^{C}) to its pathogenic isoform, PrP^{Sc}. Although three-dimensional structures of the prion proteins have been obtained by NMR spectroscopy, mechanism of the conformational change from PrP^{C} to PrP^{Sc} is still unknown and functions of PrP^{C} remain uncovered as well. In this study, a molecular dynamics (MD) codes were vectorized and parallelized, which are AMBER ver.7, AMBER ver.8 and MolTreC. To have insights into the mechanism of this transition, the biophysical properties of the recombinant protein corresponding to residues 90-231 are studied.

<Sub-Theme 5>

A molecular dynamics simulation system called PABIOS is now under development, which is designed to calculate molecular systems composed of more than a million particles on parallel computers. To perform such simulations with desired accuracy, Particle-Particle Particle-Mesh (PPPM) method is employed, which enables to compute long-range Coulomb interaction accurately. In order to improve performance of PABIOS on Earth Simulator, the algorithm for short-range interactions was intensively vectorized. A benchmark test was carried out using a RuvA-Holliday junction DNA complex consisted of 166,177 atoms. At present, PABIOS has achieved the parallelization and vectorization ratios of 55.0% and 97.5% on 15 nodes (120 processors).

<Sub-Theme 6>

Genomes of uncultured environmental microorganisms have remained mostly uncharacterized and are thought to contain a wide range of novel genes of scientific and industrial interests. A novel bioinformatics method is developed for phylogenetic classification of genomic sequence fragments derived from the environmental samples, by modifying an unsupervised neural network algorithm, Kohonen's self-organizing map (SOM). With Earth Simulator, SOMs for tetranucleotide frequencies in 210,000 5-kb sequence fragments obtained from 1,502 prokaryotes were constructed, which corresponded to all available prokaryotic sequences in public DNA databases. This SOM is used to classify 800,000 sequence fragments by Venter et al from samples of the Sargasso Sea near Bermuda.

These projects are considered to be leading edges. On the other hand, they have been developed a sort of independently and therefore enhancement of sharing information between the projects is needed for further collaborations.

6. Future Plans and Scopes

Recently, Bio and Nano Simulations are expected to produce next generation technologies for material developments in industries, which will be considered to be safe and costless for the environments. To accomplish this, computer codes are to be developed in Japan originally. This is one of the goals of Bio Molecular Simulation Consortium.

For it, it is strongly required to hybridize codes which have different functionalities using component programming. One issue is how to share data generated by the components. Recently, there are some research activities to share the dada with XML such as CML Comp (http://cml.sourceforge.net/schema/cmlComp/) in USA and AbiGrid (http://www.cineca.it/abigrid/workArea/QCMLdoc.html) in Italy. Therefore, in the consortium, it will be needed to discuss how to share the data as a future approach.

平成16年度地球シミュレータ研究プロジェクト成果報告書

コンソーシアム責任者
高田　俊和　　日本電気(株)基礎・環境研究所
著者
高田　俊和　　日本電気(株)基礎・環境研究所

1. コンソーシアムの説明

バイオシミュレーション研究者の会は、2003年3月に設立された。設立の目的は、地球シミュレータへの申請課題が、当該分野の専門家の事前評価を通じ、充分高い研究内容であることを確認し、研究者の会として、申請を行うことであった。以下で述べるように、申請課題については、会員間で相互レビューを厳格に行っており、研究内容の水準については充分注意を払っている。現在、会員数は40名を越しており、生体分子に関わる国内の研究者の多くが会員として登録されている。

2. コンソーシアムの目的

次世代の重要分野として、バイオテクノロジーが注目されている中、

1) 生体機能の発現メカニズムを、分子レベルで明らかにする。
2) それらの知見・技術を、創薬など産業上の活用に結びつける。

ことを目的として、必要な分子シミュレーション技術の構築とプログラムの自主開発を進め、それらの有効性を実証するための大型計算を地球シミュレータで行う。また、次の活動内容で述べるように、バイオ分野の地球シミュレータへの課題申請の際、その課題の研究水準が専門的見地からみても充分高いことを、コンソーシアム内でのサーキュレーションにより事前に確認し、コンソーシアムとして地球シミュレータに推挙することも、重要な活動目的のひとつである。

3. コンソーシアムの活動内容

バイオシミュレーション研究者の会の主たる活動内容は、次の二つである。第1の活動は、例年行われる地球シミュレータへの課題申請の際、その研究内容が専門的見地から見ても充分高いものであり、且つ地球シミュレータでの計算にふさわしいことを、事前審査することである。そのプロセスは、2)課題申請希望者は、決められたフォーマットに研究内容を記載し、バイオシミュレーション研究者の会の事務局に提出する、2)事務局より、全ての会員に提案書が送付され、会員のレビューを受ける、3)会員の支持を受けた研究課題のみが、地球シミュレータへサブテーマとして申請されることになる、である。このように会員同士の厳しい相互レビューの下、申請が行われている。

第2の活動は、研究者間の情報交換の場の提供である。生体分子の発現メカニズムを解明するには、ゲノム情報の解析、蛋白質のフォールディング、蛋白質の熱運動解析、電子状態の計算など多岐にわたる技術が必要である。これらの計算手法は、これまで独立な研究分野として発展してきた経緯があるが、今後はこれらの分野の技術交流の活性化による相互認識の向上を図る必要がある。申請課題の相互レビューは、この点においても充分役割を果たしていると、考えている。

4. サブグループ名

サブテーマ1：
「密度汎関数法による超大型タンパク質の全電子計算」
　佐藤文俊　東京大学生産技術研究所
サブテーマ2：
「蛋白質の高次構造変化のリアルなシミュレーション」
　斎藤　稔　弘前大学理工学部
サブテーマ3：
「第一原理からのタンパク質の折り畳みシミュレーション」
　岡本祐幸　分子科学研究所
サブテーマ4：
「正常プリオンタンパク質から異常プリオンタンパク質への構造転移プロセスの解明に関する研究」
　秋山　泰　産業技術総合研究所生命情報科学研究センター
サブテーマ5：
「分子動力学シミュレーションを用いた大規模生体超分子系の機能解析」
　石田　恒　日本原子力研究所計算科学技術推進センター
サブテーマ6：
「全ゲノム配列と全タンパク質配列の自己組織化地図作成と進化シミュレーション」
　池村淑道　総合研究大学院大学葉山高等研究センター

5. グループ間の連携

生体分子の機能発現のメカニズムを解明するのは、これまで独立した研究分野として発展してきている多方面の技術を連携させる必要がある。バイオシミュレーションの会では、このような認識の下に、1)ゲノム情報解析、2)蛋白質のab initioフォールディングシミュレーション、3)蛋白質の水溶液中での熱運動シミュレーション、4)蛋白質のつかさどる化学反応の電子状態計算、の4分野でのシミュレーションを、地球シミュレータで行っている。平成16年度、上記6サブテーマについて、研究を進めてきたが、それらは、サブテーマ6が1)に、サブテーマ3が2)に、サブテーマ2，4，5が3)に、サブテーマ1が4)に分類される。それらの研究内容の概略を、次に示す。

＜サブテーマ1＞

密度汎関数法によるタンパク質の量子化学計算ソフトウエアProteinDFを開発し、104残基の金属タンパク質シトクロムcの全電子カノニカル波動関数計算に世界で初めて成功した。このProteinDFを地球シミュレータに適用し、1,000残基規模の超大型タンパク質についての量子化学計算を実現することが、目的である。蛋白質では、極めて多くの電子がその化学現象に関与しており、このような蛋白質の全電子計算を行うことにより、分子軌道の広がりなど、従来の部分的な計算からは得られない、蛋白質全体のピクチャーが得られると考えている。

＜サブテーマ2＞

蛋白質の3次元構造を、線形に引き伸ばした状態から、計算のみから予測することは、この分野における最も難しい課題のひとつである。独自に開発したレプリカ交換分子動力学(REMD)により、アミノ酸数56個のタンパク質であるProtein G において、初めてab initioフォールディングシミュレーションに成功した。水分子の数は17,187個であった(系の全原子数は52,416)。地球シミュレータの896CPUが使われ、224個のレプリカによるレプリカ交換分子動力学シミュレーションが効率良く実行され、いろいろな2次構造が形成されるのを観測した。今後は、よりサンプリング効率の高いMUCAREMを適用していく予定である。

＜サブテーマ3＞

第1の目的は、生命を維持する上で重要な蛋白質であるヘモグロビンの立体構造変化(アロステリック効果)を、分子動力学シミュレーションプログラムCOSMOS90を用いて追跡し可視化することである。COSMOS90は、水中の蛋白質における長距離クーロン力をカットオフせずに高速にシミュレートするプログラムである。ヘモグロビンのシミュエーションを行うため、COSMOS90を地球シミュレータ上で高速に稼動するように、ベクトル化と並列化を行った。その結果、Barnes-Hut tree codeを含んだ全ての計算機能をベクトル並列化することに成功した。セットアップを行った水中のヘモグロビンについて計測したところ、0.023 sec/stepであり、米国のNAMD2.4よりも更に高速に稼動することが確認された。

＜サブテーマ4＞

狂牛病、スクレイピーなどの原因物質とされるプリオンタンパク質は、生体内で安定に存在しているが、その機能が明らかにされていないタンパク質である。生体内で正常な構造を取っているプリオンタンパク質が、感染型の異常構造を取っているプリオンタンパク質と接触することで、ミスフォールディング(折り畳み誤り)を起こして感染型の異常構造を取りアミロイド化していくことが知られているが、その感染・構造転移の機構は明らかにされていない。本プロジェクトでは、プリオンタンパク質の感染と構造の関係について、分子動力学シミュレーションから、明らかにすることを目標としている。

＜サブテーマ5＞

数百万原子のシステムを扱う大規模な並列分子動力学シミュレーションシステム(PABIOS)を開発している。PABIOSは長距離クーロン相互作用を精度良く高速に計算するParticle-Particle Particle-Mesh (PPPM)法など、最新のアルゴリズムを搭載している。166,177原子からなるRuvA-Holliday junction DNA複合体の系を用いてPABIOSのベンチマークテストを実行したところ、15ノード(120 CPU)使用時においても55.0%の並列化効率、97.5%のベクトル化率を達成することができ、計算速度も昨年度と比べて約2倍程度向上させることに成功した。

＜サブテーマ6＞

多様な地球環境由来の難培養性微生物のゲノムは、新規性の高い遺伝子を含む可能性が高く産業的に関心を集めており、ヒトの体内環境の難培養性微生物については、医薬学的にも注目を集めている。ゲノム配列解析用の一括学習型SOMは、新規性の高い配列の系統分類を可能にする、革新的なバイオインフォーマティクスである。DNAデータバンクに収録されている既知原核生物のDNA配列の全体を対象に、配列を5kbに断片化し、4連塩基の一括学習型SOMを地球シミュレータで作成した。上記の既知原核生物の配列について、25の主要系統群への分類を評価したところ、85%以上の配列が正しい系統を反映して分離していた。

バイオシミュレーション研究者の会から申請されている研究課題は、その分野のリーディングエッジに位置する研究水準を維持していると考えている。一方、このような具体的な計算手法と対象を、それぞれのサブテーマが持っているため、相互の情報交換に関しては、今後の課題として残されていると認識している。

6. 今後の計画・展望

将来のHPC (High Performance Computing)の分野における重要な領域として、マテリアルシミュレーションが注目されており、それらは、大まかに言って、ナノとバイオに分類される。何れも、次世代の産業を支える基本物質の開発に繋がるとの期待から、欧米ではプログラム開発やシミュレーションが活発に行われている。このようなマテリアルシミュレーションプログラムの多くが、アメリカ製である。期待されているように、マテリアルシミュレーションが産業競争力の強化に有効であるならば、それらのプログラムの国産化を進めなければならない。また、使い勝手の良いシミュレーション環境も用意しなければならない。

そのためには、上記サブテーマで開発されている色々な計算機能が自由に連成され、複雑な事象についてのシミュレーションが簡便に行えるようにする必要がある。その実現には、これらのプログラムをコンポーネント(部品)化して、更に、それらのコンポーネント間で授受されるデータ書式を標準化することが、有効であると考えている。

科学技術情報のXML化によるデータの標準化の試みは、米国のCMLComp (http://cml.sourceforge.net/schema/cmlComp/)やイタリアのAbiGrid (http://www.cineca.it/abigrid/workArea/QCMLdoc.html)で行われ始めている。地球シミュレータを拠点として、バイオシミュレーション関係のプログラムのコンポーネント化とデータの標準化が進むように、バイオシミュレーション研究者の会の活動内容を移行していきたいと考えている。