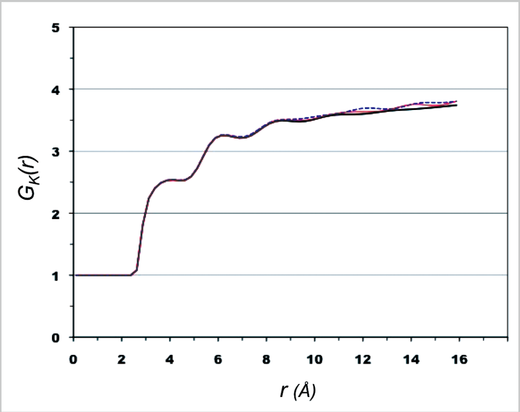
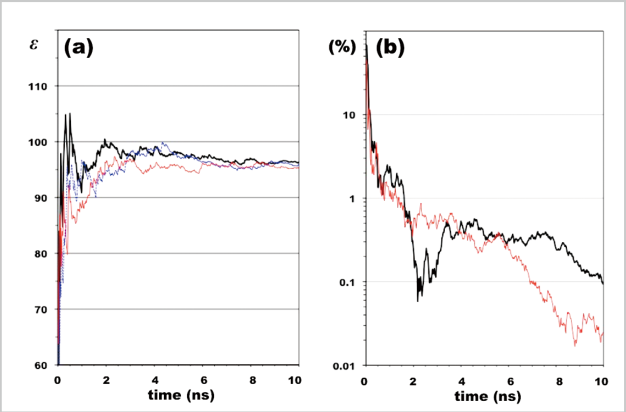
Next-Generation Integrated Simulation of Living Matter
Strategic Programs for Innovative Research Field 1 Supercomputational Life Science

BioSupercomputing Newsletter

2012.12 Vol.7



Page 6, Fig. 1 : $G_K(r)$ defined by Equation [2]. The bold black line shows the values obtained by the PME method, the thin red line shows the values obtained by the ZD method where d_c is 14 Å (α = 0), the dotted blue line shows the values obtained by the ZD method where d_c is 12 Å (α = 0). See Reference 7) for parameter α. Figure 5 in Reference 7) was modified.



Page 6, Fig. 2 : (a) Changes over time of cumulative mean dielectric permittivity obtained by Molecular Dynamics (MD) simulation. (b) Changes over time of the ratio of the second term to the first term in the curly brackets of Equation [1] expressed as a percentage on a logarithmic scale. The individual lines indicate the same items as those in Fig. 1.

CONTENTS

●SPECIAL INTERVIEW
○ Interview with "K computer" Developer regarding Efforts in Exascale and Coming Supercomputer Strategies 2-3
Executive Architect, Technical Computing Solutions Unit, Fujitsu Limited Motoi Okuda
○ Large-scale Virtual Library Optimized for Practical Use and Further Expansion into K computer 4-5
Professor, Department of Chemical System Engineering, School of Engineering, The University of Tokyo Kimito Funatsu

●Report on Research
○ Old and new subjects considered through calculations of the dielectric permittivity of water 6
Institute for Protein Research, Osaka University Haruki Nakamura
(Molecular Scale WG)
○ Development of Fluid-structure Interaction Analysis Program for Large-scale Parallel Computation 7
Advanced Center for Computing and Communication, RIKEN Kazuyasu Sugiyama
(Organ and Body Scale WG)
○ SiGN : Large-Scale Gene Network Estimation Software with a Supercomputer 8
Graduate School of Information Science and Technology, The University of Tokyo Yoshinori Tamada
(Data Analysis Fusion WG)
○ ISLiM research and development source codes to open to the public 9
Computational Science Research Program, RIKEN Eietsu Tamura

●SPECIAL INTERVIEW
○ Understanding Biomolecular Dynamics under Cellular-Environments by Large-Scale Simulation using the "K computer" 10-11
Chief Scientist, Theoretical Molecular Science Laboratory, RIKEN Advanced Science Institute Yuji Sugita
(Theme1 GL)
○ Innovative molecular dynamics drug design by taking advantage of excellent Japanese computer technology 12-13
Professor, Research Center for Advanced Science and Technology, The University of Tokyo Hideaki Fujitani
(Theme2 GL)

●Report
○ Lecture on Computational Life Sciences for New undergraduate Students 14
HPCI Program for Computational Life Sciences, RIKEN Chisa Kamada

●Promotion
○ Research System for Computational Science 15

●Event information 16

Interview with “K computer” Developer regarding Efforts in Exascale and Coming Supercomputer Strategies

Executive Architect, Technical Computing Solutions Unit, Fujitsu Limited
Motoi Okuda

Will the “K computer” class supercomputer be available onsite around the year 2015!?

—— The era of the “K computer” has begun, but it is restricted to proposal-based usage and is not yet available to everyone. This is inevitable because there is no other supercomputer that has the calculation performance of 10 PFLOPS other than the “K computer” in Japan. However, I believe many researchers dream of having an advanced supercomputer environment available on demand. “FX10”, which is a commercial model of the “K computer” has been installed inside and outside Japan, but I have not heard of 10 peta class installations. With the current situation in mind, we’d like to hear how you feel about advancements in technology. When do you believe researchers will be able to freely use 10 peta class supercomputers? How will supercomputers that exceed the “K computer” be developed, and what are the challenges?

Okuda Looking at the history of the world’s supercomputing performance competition (first place in the TOP500), the performance of supercomputers has risen approximately 1,000 times over 10 years. At this rate, it is predicted that an 1EFLOPS (exaflops) machine will be developed around the year 2018 to 2020. This is a target also held by Japan. On the other hand, in order for supercomputers to be readily available to researchers, supercomputers need to be installed all over the world and their performance may be 1/10 of that of the top supercomputer. As an actual example, at the same time the “K computer” was completed, a new supercomputer system (Oakleaf-FX) that has 1.13PFLOPS performance, which is approximately 1/10 the peak performance of the “K computer”, was installed in the Information Technology Center of the University of Tokyo. Since performance values are improved 10 times in about 3.3 years, it is predicted that a supercomputer having the performance values of today’s leading TOP500 values will be deployed throughout the world in various institutions in approximately 3.3 years. Therefore, you may foresee that supercomputers with a calculation speed equivalent to the “K computer” will be available to universities and research institutions around the year 2015.

—— If 10 peta class supercomputers become available in the year 2015, does that conversely mean a supercomputer of 100 peta class would be developed by then?

Okuda All vendors have different views and approaches. We at Fujitsu are planning to follow Japan’s national project for future product development by using our experience in the development of the “K computer”. We also will provide operation assistance and support along with application optimization support for the “K computer” which will soon be operating officially. Another mission is to continue to maintain an environment that allows further advancement of software assets, while applications for the “K computer” are developed, tuned, and optimized by researchers. Like the provision of the commercial machine “FX10”, we are developing a model to be commercialized around the year 2014 to 2015 called the “100PFLOPS Level Trans-Exa System” where “Trans-Exa” means a bridge to exascale computing. Applications developed for the “K computer” and “FX10” will run on this machine as is. The architecture will be of the same concept, and the same programming model can be applied. Of course, higher performance will be produced if the program is recompiled. A CPU in “FX10” performs 1.85 times better than the “K computer” (peak performance), and has other functions that improve operability. In “Trans-Exa”, CPU performance, network performance, and also packaging density and power consumption levels are planned to be highly improved. Although the machine will have the ability to provide up to 100 peta, we assume the number of full 100 peta scale installation would be limited. Nevertheless, onsite supercomputer environments of the several peta class should be ready for universities and research institutions around the year 2015.

The road to realizing exascale

—— What are your views on coming trends in supercomputer development?

Okuda There has been an extremely large technology spurt in this year’s TOP500. This is marked by the power efficiency (performance per power) of the first place machine, “Sequoia”, of the United States. Until this year, the “K computer” held the leading performance out of the TOP1 machines, but “Sequoia” surpassed its records with 2,000GFLOPS/KW or higher with a large margin. To state it briefly, a highly power efficient supercomputer that has never been seen before, was born. However, the performance of a core in a TOP1 machine (LINPACK performance) hasn’t changed very much over the last five years (a core is a calculation unit located in the CPU). The “K computer” has a relatively high value of approximately 15GFLOPS, but Sequoia’s values is lower and is approximately 10GFLOPS. Currently, the trend is to improve performance by increasing the number of cores rather than improving the performance of the core itself.

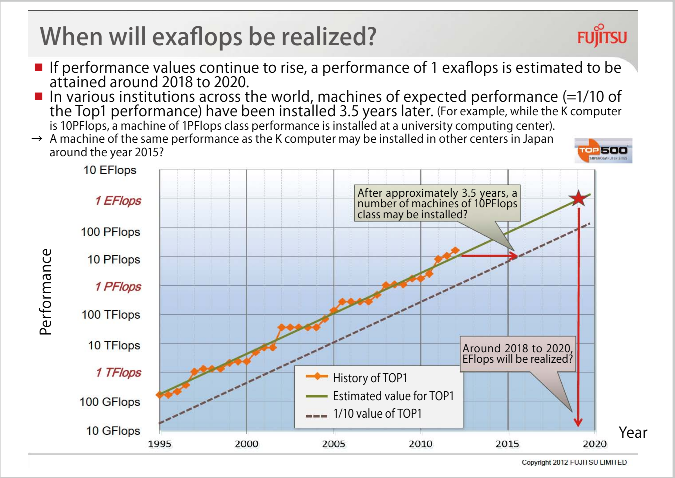
—— So I understand that lowering power consumption and increasing the number of cores are two trends in technology. In other words, could you say that they also are the challenges faced in the coming exascale development?

Okuda I agree. We are planning to realize exascale, through “Trans-Exa”, which I just explained, as the first step of technological development, and the second step will consist of research and development. The first step, “Trans-Exa” which is currently being developed, raises the performance of one core and also implements many-core processing. The “K computer” consists of 8 cores and “FX10” consists of 16 cores, and the next machine is planned to have even more cores. Along with improvements in CPU performance, we also plan to improve interconnection performance. We also are planning to lower power consumption, but this is our biggest challenge. Nevertheless, the development is moving forward with the goal of surpassing “Sequoia”. In order to attain a higher performance and density, we also plan to improve packaging density, which is the number of CPUs per dimension. We aim for exascale with these improvements which we define as the first step. During this period, we believe that there will be a big technology spurt. As predicted earlier, exascale is assumed to be achieved around the year 2018 to 2020, which is three to five years after the first step. To be honest, it is hard to predict changes in technology five years ahead. We can speculate trends for improvements in CPU calculation performance, but lowering power consumption is unclear. Technologies required for a 100PFLOPS class machine which is under development have mostly been revealed, but improvements on the current technology itself will not be sufficient to bring out 10 times the performance, and therefore we are awaiting the birth of new technologies. In addition, research and development for improving reliability will also be necessary for realizing exascale.
Parallel processing abilities will need to be improved in exasystems for application development regardless of the form it takes. It may be necessary to change the programming model. We still have not untwined the entirety of exascale. Therefore, we believe the 100PFLOPS scale system will be a platform to prepare us to advance to the future exascale. Along with progress in multi-core technology, and research and development in discovering usage of machines with SIMD capability, we hope to advance towards the future.

To develop a supercomputer that leads the world

—— So that means the machine which will become “Trans-Exa”, will be regarded as a test model for future exascale machines for engineers and researchers.

Okuda In the case of the “K computer”, application development projects



When will exaflops be realized?

- If performance values continue to rise, a performance of 1 exaflops is estimated to be attained around 2018 to 2020.
- In various institutions across the world, machines of expected performance (=1/10 of the Top1 performance) have been installed 3.5 years later. (For example, while the K computer is 10PFlops, a machine of 1PFlops class performance is installed at a university computing center).

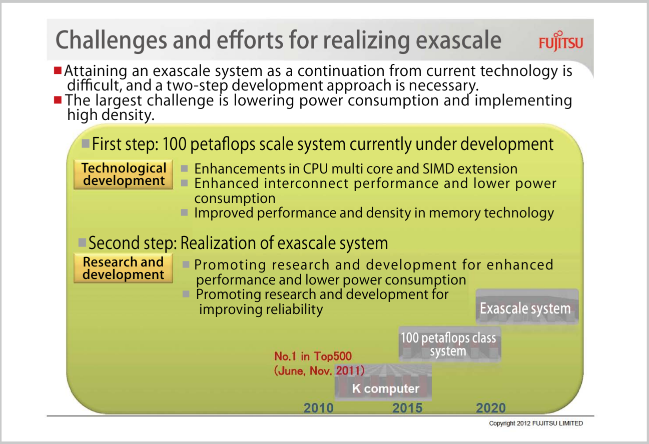
→ A machine of the same performance as the K computer may be installed in other centers in Japan around the year 2015?

After approximately 3.5 years, a number of machines of 10PFlops class may be installed?

Around 2018 to 2020, EFlops will be realized?

Copyright 2012 FUJITSU LIMITED

When will exaflops be realized?
Up to now, supercomputer performance has grown 1,000 times in 10 years, and if performance values continue to rise, a performance of 1EFLOPS is estimated to be attained around 2018 to 2020. Furthermore, supercomputers which are 1/10 the performance of the TOP1 machine in several locations in the world are likely to be upgraded to the performance of the TOP1 machine of that time in approximately 3.3 years.



Challenges and efforts for realizing exascale

- Attaining an exascale system as a continuation from current technology is difficult, and a two-step development approach is necessary.
- The largest challenge is lowering power consumption and implementing high density.

First step: 100 petaflops scale system currently under development

Technological development
- Enhancements in CPU multi core and SIMD extension
- Enhanced interconnect performance and lower power consumption
- Improved performance and density in memory technology

Second step: Realization of exascale system

Research and development
- Promoting research and development for enhanced performance and lower power consumption
- Promoting research and development for improving reliability

Copyright 2012 FUJITSU LIMITED

Challenges and efforts for realizing exascale

were initiated along with the commencement of the project. The HPCI Strategic Project started in the year 2011. What should be done using exascale should be revealed once preparation for the next step kicks off on a 100PFLOPS scale machine in the latter half of the HPCI project.

—— Yes, we're no longer in the times when calculation speed would be enhanced solely by improvements in hardware.

Okuda As the phrase "Co-Design" states, this is the age for engineers and researchers to work together to design and develop machines and applications. Like the grand challenge application developments for the "K computer", a preparation period is necessary, or else we will be missing applications to make use of the machine's performance when the exascale machine is completed. Preparation started four to five years before in various fields in the case of the "K computer", and we have now reached a stage where we can start seeing the fruit of their achievements. If positive results are available in the years 2013 and 2014, they will encourage progress for future research and development, and conversely, the development of another machine for moving onto the next step along with the "K computer" will be very important. We are developing a 100PFLOPS scale machine with this vision in mind.

—— "FX1" was released just before the "K computer", and since their architecture was similar, several research institutions and universities installed the machine early on. From your explanation, you want "Trans-Exa" to be used as a preparation for exascale in a proper manner, correct?

Okuda Yes, even if a machine with high performance is available, you will not be able to produce performance values without that level of preparation.

—— What is the largest challenge in the second step in realizing exascale?

Okuda I feel that lowering power consumption will be the largest concern. Semiconductor technologies have advanced and it is possible to increase the number of arithmetic circuits in a chip to improve CPU performance. However, if the arithmetic circuit in this high performance CPU is used for all calculations, an outrageous amount of power will be consumed, and the machine cannot be operated efficiently. Therefore, the challenge is to lower power consumption. It is difficult to develop a circuit that can calculate without using much power, but we must overcome these obstacles. Our desire and mission is to attain exascale, regardless of the unclear and unknown, through various technological developments.

—— As a developer, you cannot keep yourself from trying to aim higher.

Okuda Yes, I believe we must continue to do so. We've heard from researchers that they cannot go back to their old environment once they started using the "K computer". When the development first started, there were questions such as "Is 10 peta really necessary?" and "Are there applications that can be used? " After five years, I believe the situation has totally changed. Furthermore, Japanese researchers need some "strength" to fall back on to discuss equally with workers from across the world. People all over the world are interested in the "K computer" and are awaiting research achievements through the "K computer". Therefore, we must continue to develop a machine that can be praised on a global level.

©RIKEN

"PRIMEHPC FX10"
developed by Fujitsu
as a commercial machine

Large-scale Virtual Library Optimized for Practical Use and Further Expansion into K computer

Professor, Department of Chemical System Engineering,
School of Engineering, The University of Tokyo
Kimito Funatsu

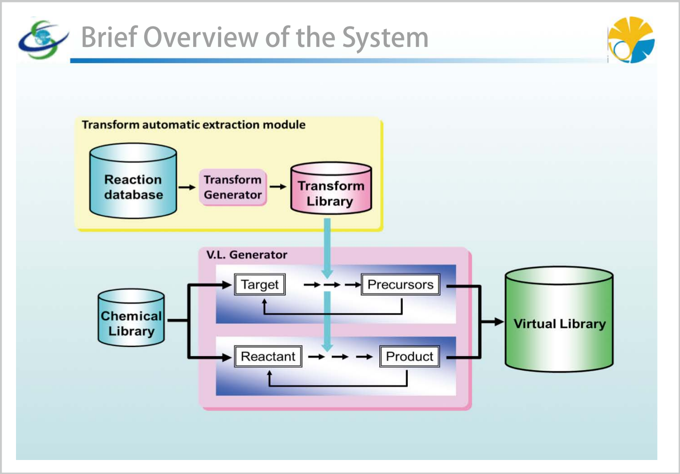
●Present status of compound library, a key for drug discovery

New drug development is very time-consuming, taking a dozen years or so. It is also said that only one out of several ten thousand compounds is sent to the medical work front. Therefore, it is a reality that both the development cost and R&D risk are extremely high. Such a new drug development starts from identification of a drug target, which is followed by discovery of a lead compound and its optimization for good activity. Then, the compound is subjected to clinical trial. The key point for success is the on-target capture of a group of lead compounds in the early stages which starts with screening of a compound library for a lead compound. This is why the potential development capability of a pharmaceutical company is dictated by the chemical variety, quality and scale of the compound library it owns.

Then what about the current status of compound libraries? It is estimated that the theoretical total of compounds which may become drug discovery targets is 10 to the sixth power. Meanwhile, existing compound libraries owned by pharmaceutical companies (megapharma) cover only several million compounds. This results in unsuccessful screening without a hit, or discovery of only low-active compounds. The large number of absent compounds has always been a major concern. For those reasons as well as improvement in the hit ratio of screening, an increase in the scale and variety of the compound library is strongly desired. From a viewpoint of searching more promising compounds exhaustively, we expect much from the use of virtual libraries constructed in computers. However, existing virtual libraries stock only several 10 million compounds at most. They are paltry compared to the theoretical total. You can get only a partial view of chemical space from their screening, so the fact is they are not very helpful. Since it is a virtual library, it has another problem. Even if we can make a short list of a group of high-score lead compounds, study of their synthesis would require great cost. If we rely only on theoretical manipulation based on a combination of atomic species and bond orders each atom can have for creation of a virtual compound structure, the long-awaited library may include many unstable compounds which cannot be synthesized.

●Feature and brief overview of large-scale virtual library

“The large-scale virtual library composed of chemical structures able to be synthesized and reaction schemes” we are engaged in constructing, is an unprecedented new virtual library that solves the aforementioned problems and develops an acceptable quality, variety and scale of the group of virtual compounds. We don’t intend to simply make a “large-scale virtual library”. It is important that the library consists of “chemical structures able to be synthesized and reaction schemes”. The library also includes synthesis routes for producing compounds. Of course, scale is also of the essence. We aim at construction of a 1 to 2 billion-scale virtual library, which exceeds the existing tens of millions-scale library, while ensuring drug-likeliness and variety. What we have developed for this purpose is a system to create new structures, in which seed structures are put into the construction system sequentially from the existing compound library covering 420 thousands compounds, followed by application of structure transformation information called Transform. Transform is structural change information on the reaction site before and after reaction, which is extracted from a reaction database. In short, Transform is a database built by extracting and storing information on changes in bond orders and structural environments at reaction sites of the reactant and the product of each reaction scheme, and or “essence of the reaction” from existing reaction databases.

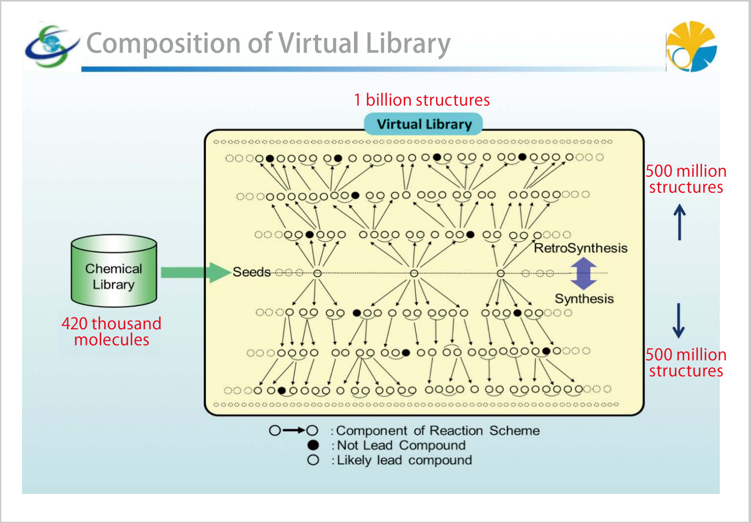


Brief overview of the large-scale virtual library system

The virtual library with ordinary synthesis routes is constructed by applying Transform information and continuously running an Ordinary Synthesis Reaction Construction System that presents product structure for reactant structure as a reaction scheme. Compound structures included in the virtual library form an ordinary synthesis tree structure that maintains a relation between reactant structure and product structure. In contrast, by continuously running a Retrosynthesis Reaction Construction System that presents a precursor structure for the target compound structure as a reaction scheme, a virtual library with retrosynthesis routes is constructed. Compound structures included in this virtual library form a retrosynthesis tree structure that maintains a relation between synthetic precursor structure and product structure. Schemes for predicting products from reactants shown in the ordinary synthesis tree domain, and a synthetic precursor with the reaction site, are proposed from the target structure in the retrosynthesis tree domain. Briefly, virtual compound structures not only from the ordinary reaction direction but also from the retrosynthesis direction suggesting the starting material, are included in the virtual library.

When a seed structure of the existing compound library is actually entered, several candidates for the product structure appear by applying Transform. Then by applying Transform to the candidates for the product structure as a reactant structure, candidates for the next-level product structure come out. On the other hand, you can track back to the starting material for a certain compound. Some compounds have small molecular size. Or they lack drug-likeliness and are considered to be inadequate as candidates for the lead compound. They are not retrieval objects of the virtual library. However, since they are necessary information in the sense of connecting synthesis routes, they are included in the library.

In this fiscal year, we are continuing scale expansion by converting the output structure into an input structure recursively and generating a multistage scheme, and aim at storing a total of 1 billion not-overlapping, unique compounds including both ordinary- and retro-reaction schemes in the virtual library as a whole. By making them branch further and adding initial seed structures, I believe a virtual library with 2 billion compounds would be attainable.



Composition of the large-scale virtual library

Arrows indicate virtual reaction scheme configuration information, open circles indicate compounds that can be a lead compound, and filled circles indicate compounds not suitable as a lead compound. Only for connecting synthesis routes, compounds with a filled circle consist of the library although they are not the target of search. In the ordinary synthesis tree domain, a product prediction scheme from the reactant is presented, and in the retrosynthesis tree domain, a proposal of a synthetic precursor for the target compound is presented.

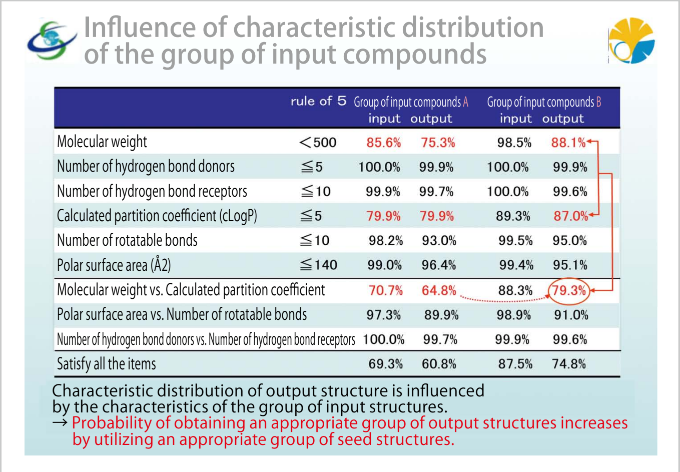
Assessment of library construction engine

While continuing further development, we make assessments for understanding the features of the virtual library construction engine and groups of output compounds. Although all compounds are not covered, when we generated 15 million compounds from 420 thousand structures used as a group of seed structures, we got 6.3 million non-overlapping, unique chemical structures. The overlapping percentage of output compounds was 58%, a little more than half. As regards novelty, when the group of 6.3 million generated compounds was compared with the existing compound library including 15 million available compounds, the overlapping percentage with commercially-available compounds was only 1.33%. Therefore, most structures output by this system were novel compounds. I conclude that novelty is fully ensured.

We are also examining the influence of character distribution of the group of input compounds. The virtual library is generated from the input seed structures. As criteria to determine whether the generated chemical structure has significance as a drug, we have ADMIT (absorption, distribution, metabolism, elimination and toxicity) characteristics. By calculating the characteristic prediction, we can assess the chemical structure preliminarily. For example, an orally administered compound does not work as drug if it is not absorbed into the body. Drugs are organic compounds and have considerably large molecular weights, so basically they are insoluble in water. Naturally, they are absorbed poorly. Since this is no good to us, those compounds need to maintain a certain level of water solubility. Solubility is not a sole determinant. A large polar surface area may increase solubility. Or, more hydrogen bond donors or receptors may help them blend with water. Like this, they can be assessed not only by solubility but also other characteristics. The empirical rule for predicting such absorption is "Lipinski's rule of five". This time, we examined the distribution of each characteristic value shown by "Lipinski's rule". As a result, we confirmed that the virtual chemical structures output by this system take over characteristics of the group of input compounds while expanding the distribution of characteristics values. We also confirmed that when using a group of compounds with high conformance ratio for each characteristic value index as a seed, the system can output a group of chemicals with high index conformance ratio. Briefly, it shows that appropriate selection of the group of seed structures enables output of virtual compounds suitable for drugs with high probability.

After this, we will actually input this large-scale virtual library to the "K computer" and have it used by users including the public. Since screening software is developed by another group, we provide the large-scale virtual library. If a library construction engine for constructing elemental chemical structures and reaction schemes is put into the "K computer", pharmaceutical companies, would-be users of the "K computer", would be able to construct a virtual library from their own compound library. The library construction engine itself has already been developed by Funatsu Laboratory. Since there is a great demand for this engine, we think we will be providing this construction engine.

We think we have almost finished preparations. From now on, we have to think how to use the large-scale virtual library for promoting its practical use toward actual drug discovery while listening to what users want. I can say that the large-scale virtual library is entering a new phase.



Influence of characteristic distribution of the group of input compounds

	rule of 5	Group of input compounds A		Group of input compounds B	
		input	output	input	output
Molecular weight	＜500	85.6%	75.3%	98.5%	88.1%
Number of hydrogen bond donors	≦5	100.0%	99.9%	100.0%	99.9%
Number of hydrogen bond receptors	≦10	99.9%	99.7%	100.0%	99.6%
Calculated partition coefficient (cLogP)	≦5	79.9%	79.9%	89.3%	87.0%
Number of rotatable bonds	≦10	98.2%	93.0%	99.5%	95.0%
Polar surface area (Å2)	≦140	99.0%	96.4%	99.4%	95.1%
Molecular weight vs. Calculated partition coefficient		70.7%	64.8%	88.3%	79.3%
Polar surface area vs. Number of rotatable bonds		97.3%	89.9%	98.9%	91.0%
Number of hydrogen bond donors vs. Number of hydrogen bond receptors		100.0%	99.7%	99.9%	99.6%
Satisfy all the items		69.3%	60.8%	87.5%	74.8%

Characteristic distribution of output structure is influenced by the characteristics of the group of input structures.
→ Probability of obtaining an appropriate group of output structures increases by utilizing an appropriate group of seed structures.

An example of influence assessment on characteristic distribution of the group of input compounds using Lipinski's rule of five as index

Old and new subjects considered through calculations of the dielectric permittivity of water

Institute for Protein Research, Osaka University
Haruki Nakamura
(Molecular Scale WG)

Textbooks say that "water has approximately 80 times as high dielectric permittivity as a vacuum, and exhibits significant polarization." Calculation of the dielectric permittivity of purified water using molecular simulation has had a long history, and some issues were clarified only recently. Looking back at these conventional and new issues might give researchers some points to keep in mind not only in a molecular simulation, but in an entire computational simulation. This paper describes an outline as follows.

The dielectric permittivity of a purified water system, ε is obtained from the statistical average of the sum total of the electric dipole moment $\{\vec{\mu}_i\}$ of individual water molecules { i } within the system, and is given as:

$$\varepsilon = 1 + 4\pi N \mu_0^2 G_K / 3k_B TV$$

where μ_0 is the electric dipole moment of a water molecule, and G_K is a scalar value called as the Kirkwood factor that is defined as:

$$G_K = \frac{1}{N}\left[\left\langle \sum_i \frac{\vec{\boldsymbol{\mu}}_i}{\mu_0} \sum_j \frac{\vec{\boldsymbol{\mu}}_j}{\mu_0} \right\rangle_{ensemble} - \left\langle \sum_i \frac{\vec{\boldsymbol{\mu}}_i}{\mu_0} \right\rangle_{ensemble}^2 \right] \qquad [1]$$

On the other hand, the distance-dependent Kirkwood factor given by the equation below serves as an important measure that reflects the alignment and structure of water molecules with extremely high sensitivity as compared to the radial distribution function (see Fig. 1 on the cover):

$$G_K(r) = \frac{1}{N}\left\langle \sum_i \left(\frac{\vec{\boldsymbol{\mu}}_i}{\mu_0} \sum_{\substack{j \\ r_{ij}<r}} \frac{\vec{\boldsymbol{\mu}}_j}{\mu_0} \right) \right\rangle_{ensemble} \qquad [2]$$

The factor shows a correlation with the electric dipole moment of water molecules on the first and second layers surrounding a water molecule, and even those on far-off layers. The second term of Equation [1] becomes zero in a sufficiently large statistical population, and therefore, $G_K(r)$ given by Equation [2] where r is infinite is equal to G_K that represents the dielectric permittivity. In other words, a value on a sufficiently far-right point on the horizontal axis in Fig. 1 corresponds to the dielectric permittivity. It is calculated as approximately 96 by the particle-mesh Ewald (PME) method, a standard calculation technique, at 300 K and 1 atm under periodic boundary conditions. The discrepancy from 80 would have arisen from the TIP3P water molecule model[1] used in the calculation.

Hydrogen-bond network relaxation in aqueous purified water is slow in terms of dynamics. Therefore, as shown in Fig. 2 (a) on the cover, simulation over a short period of time such as 1 to 2 ns is not enough for convergence of the dielectric permittivity value, and a long-time simulation at least 6 ns needs to be performed. The phenomenon, however, was pointed out in a systematic manner for the first time by Gereben and Pusztai (2011)[2]. This means that the values obtained by simulation over a short period of time in other papers published before are not reliable. My team had been faced with the same problem that the dielectric permittivity often varied in a short trajectory for several ns even using the same computational simulation method, and the finding by Gereben and Pusztai provided an unequivocal answer to our question. Furthermore, it was found that the second term of Equation [1] has not become a negligible, small amount of time average just for 1 to 2 ns as shown in Fig. 2 (b) on the cover. These findings are consequences through enhancement of computational resources.

In addition, there is another issue as regards handling long-range electrostatic forces in purified water. Only the interaction between water molecules within a certain cutoff distance d_c had often been taken into consideration because of limited computational resources although the PME method is currently used as a standard technique. A simple idea is to set d_c to as a large value as possible, but Yonetani[3] pointed out in his paper that $G_K(r)$ varied greatly by an order of magnitude in both positive and negative directions even if d_c was set to as large as 18 Å. The paper demonstrated that it not only caused a quantitative variation of the dielectric permittivity, but also failed to reproduce a qualitative water structure.

In the past, Neumann[4] proposed a method in which the artifact caused by the cutoff mentioned above is removed by the Reaction field method, assuming a dielectric material with the dielectric permittivity ε_{RF} lies outside a sphere with a radius of d_c, resulting in creation of a reaction field introduced by Fröhlich[5] . This method had been tried by many researchers, but has not been gaining high popularity recently, because it is difficult to estimate the parameter ε_{RF} in a non-uniform system such as a protein solution. Recently, Dr. Ikuo Fukuda, RIKEN, proposed the "Zero-dipole summation principle" (ZD method)[6)-8)] as one of the non-Ewald methods. This is an outstanding method that incorporates the long-range effects by imposing not only the charge neutral condition proposed by Wolf[9] but also that of electric dipole moment, allowing for high computational accuracy with a simple algorithm. By using the method, almost identical values of dielectric permittivity and $G_K(r)$ as those from the PME method can be obtained as shown in Figs. 1 and 2, even though d_c is 12 Å to 14 Å. Namely, the ZD method works well even in a state where molecular interactions are cut off within a short range and only short-range interactions are taken into consideration. Interestingly, the ZD method gives exactly the same equation under certain conditions as that in the Reaction field method where ε_{RF} approaches infinity. It was also found that the ZD method has some characteristics in common with other various non-Ewald methods proposed recently[8]. It seems that common physics works as a basis in these methods in that no periodic boundary conditions are given.

It is often the case in computational science that researchers compete with each other on computational speed according to given algorithms. This is definitely necessary, but creation of the original algorithms or models would open up completely new possibilities. My team has been working on studies that apply the ZD method mentioned above to heterogeneous systems such as protein and DNA solutions not as a periodic boundary system, but as a three-dimensional torus system, and are beginning to produce favorable results.

In computational research using new algorithms and models, researchers always conflict with reviewers who are insistent on the "Ptolemaic theory" and have some difficulties in publishing a paper. However, successful research results have a significant ripple effect. In fact, eliminating the use of the periodic boundary system in computation will allow us for easier high-speed simulation of many biological supramolecules with fewer computational resources.

Finally, my sincere acknowledgements are for the collaborators of this research, Dr. Ikuo Fukuda (RIKEN), Dr. Narutoshi Kamiya (Institute for Protein Research, Osaka University) and Dr. Yasushige Yonezawa (Kinki University).

[References]
1) W. L. Jorgensen et al., *J. Chem. Phys*. **79**, 926 (1983); 2) O. Gereben, L. Pusztai, *Chem. Phys. Lett.* **507**, 80 (2011); 3) Y. Yonetani, *J. Chem. Phys.* **124**, 204501 (2006); 4) M. Neumann, *Mol. Phys.* **50**, 841 (1983); 5) H. Fröhlich, "Theory of Dielectrics" Clarendon Press (1958); 6) I. Fukuda et al., *J. Chem. Phys.* **134**, 164107 (2011); 7) I. Fukuda et al., *J. Chem. Phys.* **137**, 054314 (2012); 8) I. Fukuda, H. Nakamura, *Biophys. Rev.* **4**, 161 (2012); 9) D. Wolf et al., *J. Chem. Phys.* **110**, 8254 (1999)

Development of Fluid-structure Interaction Analysis Program for Large-scale Parallel Computation

Advanced Center for Computing and Communication, RIKEN
School of Engineering, The University of Tokyo (until September 30, 2012)
Kazuyasu Sugiyama
(Organ and Body Scale WG)

Blood flow has the functions of maintaining healthy conditions (such as hemostasis, transport of substances, removal of foreign bodies and body temperature control). For example, if a blood vessel wall is damaged, platelets adhere to the wall and form thrombi, allowing for repair of the vessel. On the other hand, overgrown thrombi caused by any factors produce vessel occlusion, resulting in cardiac or cerebrovascular diseases with a high risk of sequelae or death. Accurate prediction of normal and abnormal blood conditions by a physical computation methodology based on logical rationales contributes to advancement of treatment and drug discovery. Our research group developed an interaction analysis program that combines fluid and structure dynamic actions (the ZZ-EFSI code) with a focus on blood flow phenomena at the continuum level.

Blood contains massive amounts of blood cells that deform flexibly. In a microcirculation system with a submilimetric scale vessel, the deformability of red blood cells and the characteristics of dense particulate flow dictate blood flow functions. The laws and the principles of dynamics (the conservation law, and constitutive equations that describe viscosity and elasticity) of blood and blood cells are simple. But, the blood system involves phenomena over a wide range of the time-space domain and exhibits complicated behavior. We have developed the ZZ-EFSI code with the aim of performing large-scale computation based on simple laws and principles. To this end, we determined new equations to be implemented and reexamined the computation scheme and algorithms, without having tuned conventional interaction analysis codes, keeping in mind that we should exploit the performance of the "K computer" [1, 2].

The recent scalar-type supercomputers such as the "K computer" are characterized by hierarchical parallel processing (MPI parallel computing accompanied by communication between compute nodes, thread-level parallelism between cores within compute nodes, and multiplex operation within cores). We developed an Eulerian method that does not require the process of mesh generation and reconstruction (more specifically, all physical quantities are updated on a spatially-fixed mesh). In this computing program, the rectangular computational domain is divided into meshes in the X-, Y- and Z-axis directions to describe equations and allow for MPI domain decomposition. The Eulerian method is compatible with any hardware structure at all layers and excels in expanding the scale of computation. Standard fluid applications tend to access memory too frequently with respect to the operation load. This means that a huge backlog is caused frequently in a scalar-type supercomputer that takes more time to read and write memory as compared to performing operations. The efficiency (the effective computation speed as compared to the theoretical peak performance) of the "K computer" is limited to about 10 percent. To address this issue, we developed an algorithm that requires less frequent memory access to improve the computation speed. Fig. 1 shows the performance of fluid-structure interaction analysis by the "K computer". The efficiency as a single node was 46.6 percent, a sufficiently-high level for continuum dynamics computation by a scalar-type supercomputer. In addition, small changes in the efficiency with increase in the number of compute nodes mean high linear scalability. We performed computation of the system that contained approximately 5 million dispersed objects by using 82,944 nodes with 6.96 × 10^{11} grid points, and successfully achieved an effective computation speed of 4.54 petaflops.

We have simulated a cerebral arteriole flow that included red blood cells and platelets (see Fig. 2 (a)). Fig. 2 (b) and (c) shows the path of some platelets. Under the condition that no red blood cells were included (Fig. 2 (b)), individual platelets had small changes in radial coordinates and moved along the vessel axis almost in a straight line. On the other hand, under the condition that red blood cells were included (Fig. 2 (c)), platelets had larger changes in radial coordinates, or in other words, dispersed easily. These results mean that red blood cells disrupt the blood flow and cause massive fluctuation of platelets, resulting in increased opportunities for platelets to approach the vessel wall. This theory is consistent with experimental results that suggest the importance of disruption effects by red blood cells against formation of platelet thrombi.

We have introduced a model of platelet adherence to a damaged vessel wall [2] and are working on a study that demonstrates experimental knowledge concerning thrombus formation. If the efficacy of drugs can be assessed based on patient-specific data in the future, it will facilitate development and discovery of new, attractive drugs. Toward this goal, we will be committed to creating models of changes in physical properties, the process of coagulation and dissolution, and biochemical reactions.

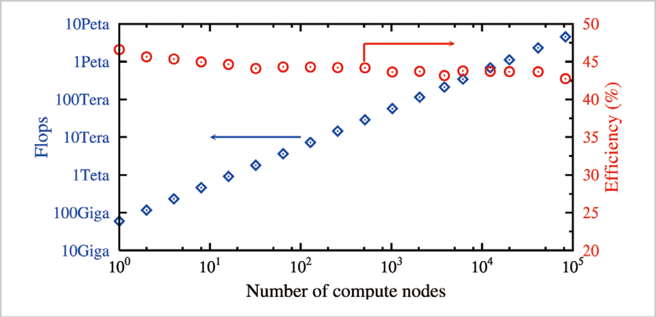


Fig. 1 : Weak scaling parallel performance of the ZZ-EFSI code in the "K computer" (left: effective computation speed, right: efficiency rate). The number of grid points per compute node is 512 × 128 × 128.

[References]
[1] BioSupercomputing Newsletter, Vol. 2, p. 11.
[2] BioSupercomputing Newsletter, Vol. 6, pp. 2-3.

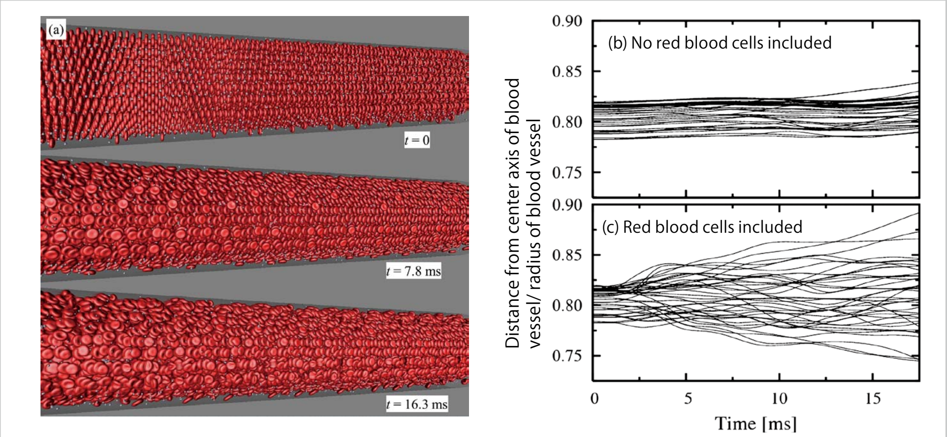


Fig. 2 : Computation results of massive dispersed objects in a blood vessel with a diameter of approximately 100 μm. (a) Snapshots of blood cell distribution (red: red blood cells, white: platelets). Blood flows from left to right. (b) and (c) Time-varying changes of radial coordinates of platelets (meaning the distance from the center axis of the blood vessel). The path of platelets is different depending on the presence or absence of red blood cells.

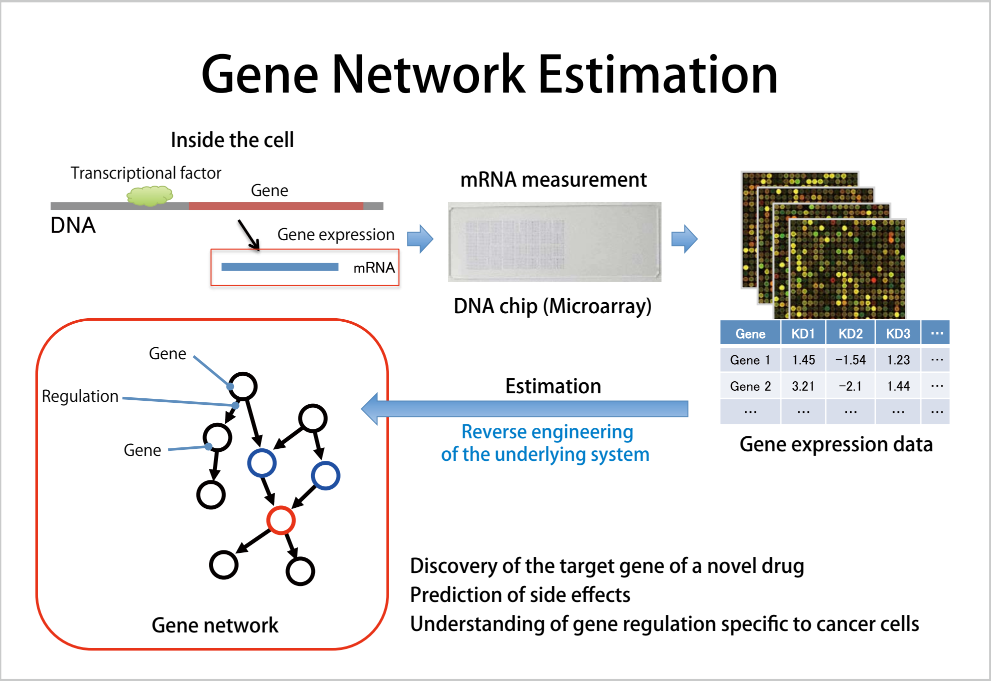
SiGN: Large-Scale Gene Network Estimation Software with a Supercomputer

Graduate School of Information Science and Technology,
The University of Tokyo
Yoshinori Tamada
(Data Analysis Fusion WG)

Human cells are said to have about 20 to 30 thousand genes. The human body is mostly composed of proteins. The gene is a blueprint for proteins made in the cell. As well as the kind of protein, the timing and quantity of protein production is also regulated by special genes. Those genes (≒proteins) are also regulated by another gene. Briefly, genes form a complicated regulation network. Most of the system is still poorly understood. Even among the same human cells, they have different networks in different organs. Drugs modify gene networks and cancer cells have destroyed networks. Gene network estimation is an approach to infer or estimate such gene regulatory networks (=gene network) from measurable data by mathematical, statistical and informational scientific methods. Although it is impossible to measure all proteins produced in the cell with current technologies, the amount of mRNA synthesized prior to protein production is measurable for every gene. Data measured like this is called gene expression data. Data obtained from one measurement are a snapshot of cell status. It is impossible to infer or estimate regulations between genes only with this one measurement. Massive data are necessary for that. Therefore, we collect data necessary for estimation by applying various stimuli to the cell, collecting cells from patients with a particular disease or taking data temporally at regular time intervals. Inference and estimation of gene networks enables clarification of regulations between genes by exhaustive computation instead of the conventional time-consuming method of searching for genes one by one and repeating experiments. It is expected that this approach will enable efficient development of new drugs, identification of cancer-specific genes, and understanding of the functions of such genes.

SiGN is software for estimating a gene network with a supercomputer from gene expression data. As the gene network, various models have been proposed. However, every model has both merits and demerits. None of them is by far the best. After deciding a model, we still have to choose a method for estimating parameters from the data. Those methods also have good and bad points. SiGN is a gene network estimation software implementing multiple gene network models and estimation algorithms, both of which requires vast amount of computation time assuming computation using a supercomputer. In particular, SiGN is composed of three sub-programs, SiGN-BN using static and dynamic Bayesian networks, SiGN-SSM using a State Space Model (SSM) and SiGN-L1 implementing a parameter estimation method by L1 regularization. SiGN-BN implements a new algorithm called NNSR. Conventionally, gene network estimation using Bayesian networks was applicable to about 1000 genes. Now it is applicable to all genomes (all genes) thanks to NNSR. Temporal data allows SiGN-SSM to estimate dynamic gene networks that are able to be simulated. It does not give the network structure but the strength of relationships among all genes as mathematical values. Thanks to supercomputers, network structures which have been difficult to compute, are now computable with a degree of confidence. L1 regularization was originally applicable to large-scale gene networks. However, the computation time of conventional methods is not enough to estimate networks focusing on individual differences in gene expression. By using the K computer, it is able to be computed within a realistic time-frame.

Development of SiGN is targeted mainly at the K computer and Shirokane, a supercomputer of the Human Genome Center. Several sub-programs have already been installed in Shirokane and are available for users. For more information, please contact the SiGN website at http://sign.hgc.jp.



Gene	KD1	KD2	KD3	…
Gene 1	1.45	−1.54	1.23	…
Gene 2	3.21	−2.1	1.44	…
…	…	…	…	…

Discovery of the target gene of a novel drug
Prediction of side effects
Understanding of gene regulation specific to cancer cells

ISLiM research and development source codes to open to the public

Computational Science Research Program, RIKEN
Eietsu Tamura

1. A software research and development project unprecedented anywhere in the world

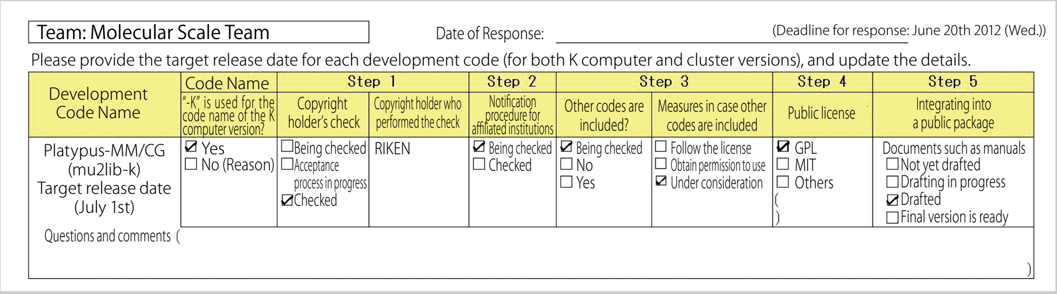
One of key goals of the ISLiM project is to develop software that allows the K computer to realize its ability, to publish excellent scientific papers, and to make the software available to the K computer.
One of key advantages of ISLiM is that it consists of about thirty codes, which comprehensively cover everything from the molecular scale to body scale, and from simulation to data analysis, and that they are tuned for the K computer in a sophisticated way. The software built in this comprehensive manner in the life science/healthcare field is unprecedented in the world, and confers significant value as a Japanese software asset used in research applications as well as educational applications.

2. Activities for releasing the source code

Since the latter half of 2010, this project, in collaboration with "Research Group for Promoting the Utilization of Next-Generation Supercomputers in the Drug Discovery Industry," has promoted exchange of information with the domestic pharmaceutical industry in terms of how to utilize the software after its completion. In these discussions, we reaffirmed that when publicly releasing such leading-edge software, it is important to publicly release the source code that can be verified and corrected by users rather than to provide binary code such as commercial software that is much used, and can be provided together with the structure for providing prompt support. In addition, we reaffirmed that there is a need not only for the "K computer" version, but also for a "cluster system version" commonly used in business.

● Sharing goals via principal developer's meetings

To release the source code, it is important that software developers specifically understand the significance of and the required process for releasing the source code, and dispel doubts and concerns about the release. The ISLiM project was established and discussed by principal developer's meetings three times, i.e., on November 9th 2011, on February 21st 2012, and on July 23rd 2012. In these meetings, we explained and discussed the purpose as well as received useful advice from Mr. Takahiro Honma, a specially appointed professor of the Center for Industrial and Governmental Relations of The University of Electro-Communications, and Kazuki Shigemori, a patent attorney of Andersen Mori & Tomotsune, who has good knowledge of intellectual property with respect to software. In terms of promotion of the project, we provide a "Flow Chart to Prepare for Releasing ISLiM Developed Software" so that the persons responsible for program development can understand the standard process of the release, and share in the progress as shown in Fig. 1.



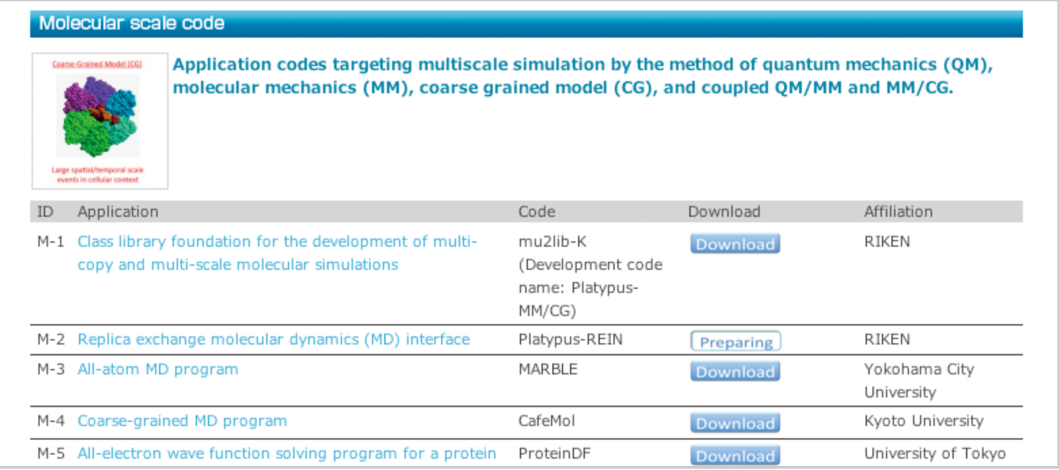
Team: Molecular Scale Team   Date of Response: ____________ (Deadline for response: June 20th 2012 (Wed.))

Please provide the target release date for each development code (for both K computer and cluster versions), and update the details.

Development Code Name	Code Name	Step 1		Step 2	Step 3		Step 4	Step 5
	"-K" is used for the code name of the K computer version?	Copyright holder's check	Copyright holder who performed the check	Notification procedure for affiliated institutions	Other codes are included?	Measures in case other codes are included	Public license	Integrating into a public package
Platypus-MM/CG (mu2lib-k) Target release date (July 1st)	☑ Yes ☐ No (Reason)	☐Being checked ☐Acceptance process in progress ☑Checked	RIKEN	☑ Being checked ☐ Checked	☑ Being checked ☐ No ☐ Yes	☐ Follow the license ☐ Obtain permission to use ☑ Under consideration	☑ GPL ☐ MIT ☐ Others ( )	Documents such as manuals ☐Not yet drafted ☐Drafting in progress ☑Drafted ☐Final version is ready

Questions and comments ( )

Fig.1 : Sharing the progress status



Molecular scale code

Coarse-Grained Model (CG)
Large spatial/temporal scale events in cellular context

Application codes targeting multiscale simulation by the method of quantum mechanics (QM), molecular mechanics (MM), coarse grained model (CG), and coupled QM/MM and MM/CG.

ID	Application	Code	Download	Affiliation
M-1	Class library foundation for the development of multi-copy and multi-scale molecular simulations	mu2lib-K (Development code name: Platypus-MM/CG)	Download	RIKEN
M-2	Replica exchange molecular dynamics (MD) interface	Platypus-REIN	Preparing	RIKEN
M-3	All-atom MD program	MARBLE	Download	Yokohama City University
M-4	Coarse-grained MD program	CafeMol	Download	Kyoto University
M-5	All-electron wave function solving program for a protein	ProteinDF	Download	University of Tokyo

Fig.2 : A part of the download site (http://www.islim.org/islim-dl_e.html)

● Each software program available from the download site as it is ready to be released

We have set up a download site (http://www.islim.org/islim-dl_e.html) in 2011 so that our source code is widely utilized not only in academia but also in industry, and released each software program when it was ready to be released as shown in Fig. 2. A goal set by principal developer's meetings is to release 50% of all the software in April 2012, and 100% in October 2012, six months before completion of the project. We plan to conduct promotional activities, such as debriefing meetings and training sessions, during these six months.

3. Current status of the public release of the source code, and challenges for the future

Thirty four (34) codes were involved when we started research and development in 2006. Some software programs have been consolidated into one program in the final stage of development and, as a result, about thirty (30) codes are expected to be provided. For the latest information on the release status, visit our download site. Access to either the "K computer" version or the "cluster system version" can be switched over by a dedicated compiler.
We have been investing a great deal of resources to research and development of the software for six years. Challenges that remain to be addressed are how we bring such valuable Japanese software asset to next players effectively by a project completion.

Understanding Biomolecular Dynamics under Cellular-Environments by Large-Scale Simulation using the "K computer"

Chief Scientist, Theoretical Molecular Science Laboratory,
RIKEN Advanced Science Institute
Yuji Sugita
(Theme1 GL)

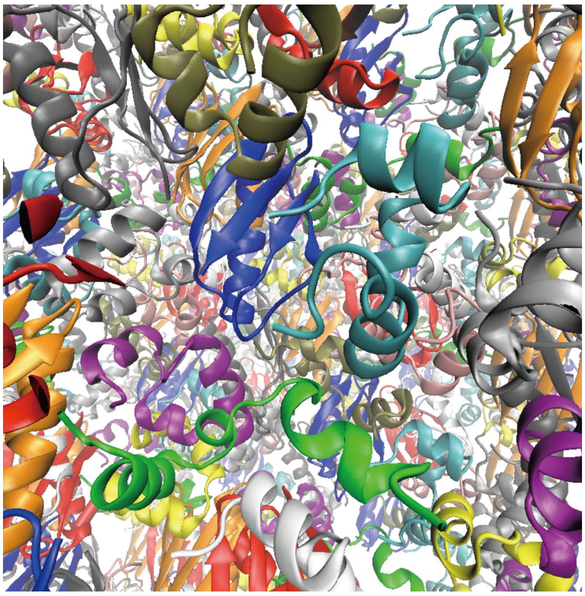
Overview of the Project

Simulation studies in life science are now one of the most active research fields with rapidly progressing development of methodologies and algorithms. Many experimental data such as genome information, atomic structures and intracellular protein expression information are obtained one after another. Conventional life science focused on only the experimental data. However, it's about time that we integrate these experimental data and promote better understanding of biological phenomena in a living system. In that sense, simulations achieved by the excellent computing capability of the "K computer" can be said to play a great role in changing life science into a new, predictable and controllable research system. It was in these circumstances that Strategic Programs for Innovative Research Field 1 Supercomputational Life Science began. In the theme we are engaged in (biomolecular simulations under cellular-environments), we aim at understanding and prediction of intracellular molecular dynamics by performing large-scale computer simulations on a molecular and cellular scale in which we focus intensively on the cellular environment. Many biomolecular simulations have been done so far. However, most of them tried to reveal the behavior of a single protein or DNA in aqueous solution or in a lipid bilayer membrane. When computed from the number of proteins in the cell and size of the cell, there is no doubt that the cytoplasmic environment is very different from the environment of aqueous solution. This has also been revealed experimentally. Just near one acting protein, there are other proteins. It is not yet completely understood how such an environment (intracellular molecular crowding environment) influences protein structure, stability and functions. In previous theoretical studies, the macromolecular crowding was studied using a very simple model in which a protein molecule was approximated to one particle. But, under this assumption, detailed intra-molecular interactions cannot be taken into account in the simulations. In this project, we would like to look at a lot of proteins under conditions close to the cellular environment on an unprecedented scale.

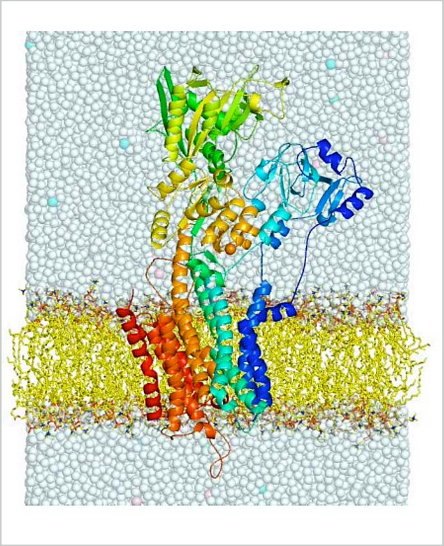
We have the following three sub-projects: The first is "Molecular and cellular-scale simulations in the cellular environments" for developing a new research field by cooperation between single-particle simulations and molecular dynamics simulations. The second is "Substance transport across the cell membrane by membrane proteins" for achieving a quantitative and predictive molecular simulation by free energy calculations based on extensive molecular dynamics simulations. And the third is "Understanding the dynamic structures and functions of DNA and proteins in a nucleus", analyzed with an all-atom and coarse graining molecular dynamics calculation. For understanding life phenomena and linking them to prediction, it is essential to understand a "whole cell" by integrating molecular information with cellular information. For this purpose, at the end of the project, we would like to define the next challenge aiming at "Whole cell simulation" by using the findings obtained from research and development.

Goal of Molecular Simulation Studies

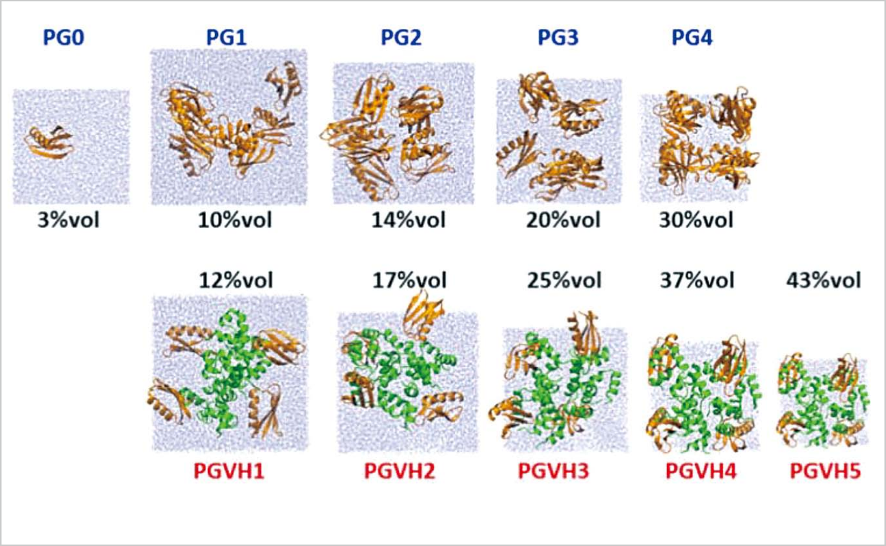
In order to realize Theme 1, "Molecular and cellular-scale simulations in the cellular environments," two elements are necessary. One is simulation of the "considerably slow motion" of biomolecules such as proteins and nucleic acids. The other is elucidation of cellular functions from the molecular point of view by connecting molecular-scale (in the atomic detail) studies with cellular-scale (in the molecular details) studies. As for the former, we would like to attempt the simulation of millisecond-scale structural changes. Although I said it is slow, it is slow as our researchers' standards. Actually it is an extremely fast molecular movement. However, in current molecular simulation studies, the time scale of long dynamics is micro-seconds. Therefore, millisecond molecular movement is a thousand times longer than microsecond molecular movement, and thus far slower. The computation environment available to us before completion of the "K computer" was about 100TFLOPS or so. Thanks to the advent of the "K



Understanding of the cellular environments and their role on biomolecules is the first step toward the whole cell modeling.



In order to see the molecular dynamics of an ion pump in biological membranes by simulation, a long stream of computing including about 260 thousand atoms such as proteins, lipid bilayer, water and ions, as well as their molecular interactions, is required.



Simulation by changing protein concentration. Protein molecules crowded in cytoplasm are simulated by molecular simulation, showing the differences in hydration effect between in dilute condition and in crowded environment.

computer", it is now enhanced 100 times (the calculation performance of "K computer" is about 10PFLOPS (10,000TFLOPS)). It is actually considerably difficult to make a calculation 1,000 times longer by similar utilization of computers. Therefore, by developing novel computational approaches, we try to see millisecond dynamics. For this purpose, we have already been developing advanced parallelization techniques. For example, we have an algorithm called the multi-dimension replica-exchange molecular dynamics calculation. In this method, molecular dynamics calculations at different temperatures or with different parameters are performed in parallel for the copy of a system called a replica, and by exchanging temperature or another parameter at a certain frequency, the calculation can be accelerated. It is difficult to do parallelization with several ten thousand CPUs for a single molecular dynamics calculation. However, in the replica-exchange method, the molecular dynamics calculation of each replica is parallelized with several hundred to several thousand CPUs. By providing several ten to several hundred replicas, simultaneous and efficient utilization of several ten thousand CPUs is enabled. We have to integrate those various methods. Although we use the "K computer" having unprecedented computing performance, what we can do only by computation is limited as I have already described. Therefore, I believe collaboration with experiment is of the essence in order to implement this project. For example, simulations can collaborate with structural biology, in particular, X-ray crystallography and NMR (nuclear magnetic resonance). In addition, we can collaborate with single-molecular experiments to measure dynamics of proteins and protein complexes in cell in a most direct way. Experimental difficulties often arise from the time-resolution, whereas simulations usually suffer from insufficient sampling due to the short computational time. At milliseconds, real-time experimental measurements correspond with the simulation results at the same time scale.

●Results of Molecular Dynamics Simulations

While proceeding with development and upgrading of software for efficient utilization of the computing performance of the "K computer", some simulation studies have already been implemented and produced results. One of them is the simulation, which takes account of the cytoplasmic molecular crowding environment, and we are taking the lead in implementing this simulation study. In the study, we could show the differences in hydration of proteins between in dilute condition and in crowded environment. This difference is essentially important for protein conformational stability, dynamics, and biochemical functions. Since microscopic parameters for the intracellular environment are difficult to measure experimentally, this computational result can be said to be quite helpful also for experimentalists. By utilizing the "K computer" hereon, molecular crowding analysis in much large-scale systems will be studied. Further progress in this research is expected. Molecular simulation of membrane transport protein has also made great strides. Conventionally, it took some time to provide conformational dynamics of membrane proteins by molecular simulation after clarification of X-ray crystal structure. Recently, we can perform molecular dynamics calculations including a lipid bilayer membrane immediately after the determination of the X-ray crystal structure. Structural change taking place partly in the transport cycle of membrane transport proteins has already been unveiled by simulations and experiments. We expect to perform molecular dynamics simulations to reveal substance transportation across membranes by membrane transport proteins using the full power of the "K computer". The project we are engaged in is expected to contribute not only to basic science, but also drug discovery and medical care. The information observed here would be useful for such applications in near future. Finally, I would like to emphasize again the importance in the developments of simulation methodologies and models to achieve our research goals. We also need strong computational skills to realize large-scale molecular and cellular simulations in this project. For this, the participation of young researchers or students is requisite. To increase the number of those people is also one of the tasks of the project over the long term.

Innovative molecular dynamics drug design by taking advantage of excellent Japanese computer technology

Professor, Research Center for Advanced Science and Technology,
The University of Tokyo
Hideaki Fujitani
(Theme2 GL)

●Shooting for Real Drug Discovery

Thanks to the advent of the K computer, improvement in computer capacity enabled molecular dynamics calculation to see how a drug molecule acts on and binds to a disease target protein. With this step, an IT innovation in drug discovery, in which atomic-level changes in protein shape are revealed for drug design, is about to begin.

In Strategic Programs for Innovative Research Field 1 Supercomputational Life Science, Theme 2 Simulation Applicable to Drug Design which we are engaged in, we make the fullest possible use of the K computer to establish new Computer Aided Drug Design (CADD) technologies for innovating the drug discovery process, as well as to discover real drugs. Although they are both within the field of molecular simulation, Theme 1 deals with broader phenomena with biological importance. In Theme 2, objects are narrowed down to disease target proteins which are the target of drug discovery. This is the feature of Theme 2.

Until now, neither national institutes nor national universities have discovered drugs independently. This is because drug development including clinical trials is very expensive, costing 20 to 30 billion yen. In addition, in the step before clinical trials, most institutions and universities unfortunately have facilities that are not suitable to synthesize numerous drug candidate compounds. Since antibody drugs are proteins anyway, basically any university can synthesize them and does not have to rely on pharmaceutical companies for synthesis. However, synthesis of low-molecular weight compounds costs a fortune for their design. Therefore, it is impossible at present for universities and institutes to do the entire process of drug discovery from the development phase to clinical trials.

For those reasons, we arrange tie-ups with pharmaceutical companies from the very beginning in our project, and proceed with the development in the form of collaborative study. We are obliged to get pharmaceutical companies involved to translate our project into real drug discovery. You may think government funds are being used for business, but the facts are contrary to that. In the research and development phase, companies cannot avoid expenses. In order to get them involved in collaborative studies even in this situation, we have to produce convincing simulation results. For discovering a real drug, both we and pharmaceutical companies have to take risks and make a serious effort.

●It is Important to Become a Pioneer of IT Drug Discovery

In these twenty years, drug discovery by computer simulation has been attempted many times but not yet fulfilled. The major reason for this is we did not have enough computation power to calculate the protein itself. In order to simulate a phenomenon in which a compound that may become a drug binds to a protein in solution and inhibits the function of the protein, at least 50 to 100 thousands of atoms are involved and 200 thousands of atoms are involved when the protein is large. Then about several million calculations are required. In addition, it was recently revealed that whether a compound is active is not is determined by simulating only one protein molecule, and simulation of the whole system is necessary. Then, the number of atoms involved easily surpasses 1 million. At last, the K computer has given us the computation power that we wanted. So far it has been determined empirically whether a compound binds to a target protein, but, now it becomes clear by computer-aided calculation. At last we have created environments for developing drugs logically.

The IT drug discovery (Fig. 1) efforts we are engaged also started in Europe and the US at almost the same time. You may know Mr. David E. Shaw (US) who made a special computer for MD calculation, ANTON, and is proceeding with IT drug discovery together with Mega Pharma (mega pharma company). Due to those activities, Japanese pharmaceutical companies which expressed their doubts became interested in IT drug discovery. However, since it has not yet produced any favorable results worldwide, they have not reached the stage for investing in and working on IT drug discovery. So we decided to become a pioneer in this field.

Now that we are on the starting line with other developed countries, it is important to start off as a pioneer. While collaborating with researchers from pharmaceutical companies in efforts geared toward real drug discovery and studying what to calculate, which result we should use and how to use the result in compound design, I believe we can expand the horizons of research and development of IT drug discovery. After they understand that they really can develop drugs through IT drug discovery, supercomputers comparable to the K computer will be introduced in major pharmaceutical companies possibly 5 years from now to promote further research. Opening up such a new epoch is the final goal of this project in a sense, and I also think it is one of the important missions for the K computer.

●High-accuracy Prediction of Drug Efficacy by Use of the High Computing Power of the K computer

Since many drugs target proteins, we have to find a compound that interacts strongly with the intravital target protein (ligand) to discover a more effective drug. We aim at rapid and efficient development of low molecular weight drugs in the following manner. By use of molecular dynamics calculations using the supercomputer, we run a simulation of a system including a target protein and a drug candidate compound, investigate protein-compound interaction, and design a new compound which acts only on the target protein.

For this purpose, we devised the MP-CAFEE method (Fig. 2) which is an algorithm for calculating free binding energy using the Jarzynski relation for free energy difference and non-equilibrium work discovered by Jarzynski in 1997. In this method, molecular dynamics calculations are performed on multiple intermediate states between two states. In one state, full interaction exists between the compound and other molecules. In the other virtual state, interaction disappears completely and the compound is uncoupled. Then, the free binding energy is calculated from the work necessary for the transition to the adjacent state. This enables accurate computation of the free binding energy between the target protein and the compound in solution. The feature of the MP-CAFEE method is that its accuracy is always high for any kind of protein and is not influenced by protein characteristics due to atomic-level calculation. Since such a huge calculation power is

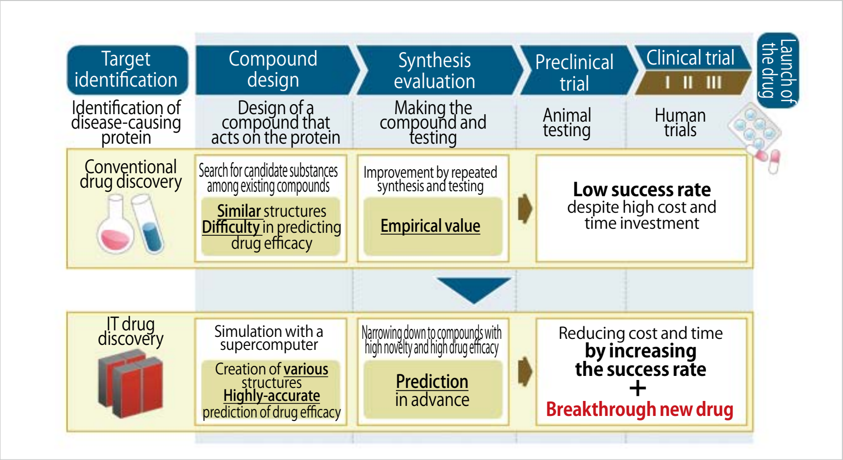
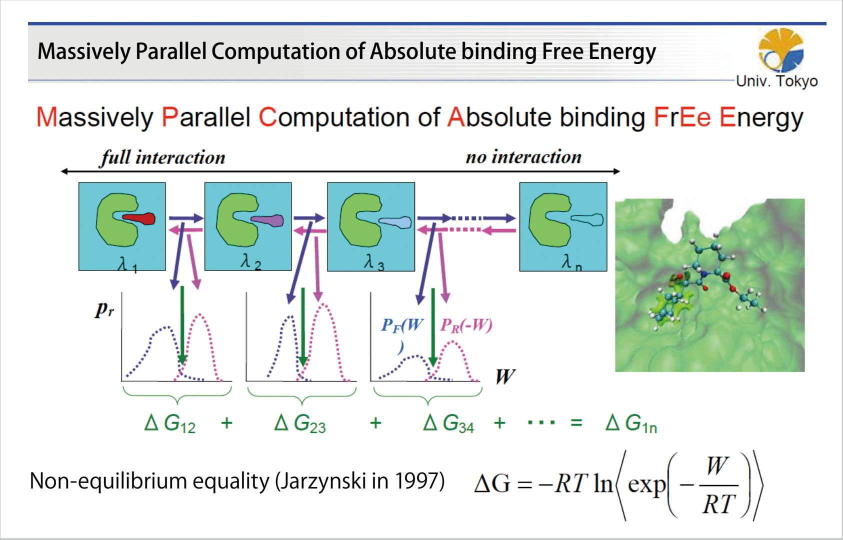

Target identification
Compound design
Synthesis evaluation
Preclinical trial
Clinical trial
I II III
Launch of the drug
Identification of disease-causing protein
Design of a compound that acts on the protein
Making the compound and testing
Animal testing
Human trials
Conventional drug discovery
Search for candidate substances among existing compounds
Similar structures
Difficulty in predicting drug efficacy
Improvement by repeated synthesis and testing
Empirical value
Low success rate
despite high cost and time investment
IT drug discovery
Simulation with a supercomputer
Creation of various structures
Highly-accurate prediction of drug efficacy
Narrowing down to compounds with high novelty and high drug efficacy
Prediction in advance
Reducing cost and time
by increasing the success rate
+
Breakthrough new drug

Fig.1 : Difference between conventional drug discovery and IT drug discovery
Figure by Fujitsu Limited


Massively Parallel Computation of Absolute binding Free Energy

Massively Parallel Computation of Absolute binding FrEe Energy

$$\Delta G_{12} + \Delta G_{23} + \Delta G_{34} + \cdots = \Delta G_{1n}$$

Non-equilibrium equality (Jarzynski in 1997) $\Delta G = -RT \ln\left\langle \exp\left(-\frac{W}{RT}\right)\right\rangle$

Fig.2 : Massively parallel computation of absolute binding free energy MP-CAFEE

In MP-CAFEE, massive molecular dynamics calculations are performed while changing drug-interaction parameters with proteins and a compound (drug) in a state of thermal agitation in solution. The binding free energy of the drug with the target protein is calculated from the work distribution when the parameter changes.

required for that, people often asked to us what kind of computer on earth we were going to use when we announced the project in 2005. However, thanks to the development of the K computer, conditions for producing tangible results are now satisfied.

In May 2011, we began work on porting Massively Parallel Computation of Absolute binding Free Energy (MP-CAFEE) into the K computer. So far, we have confirmed that MP-CAFEE works correctly in the K computer. Presently we are making adjustment for faster calculation. In August, we started calculations for actual development of drugs against target proteins of cancer and leukemia by using MP-CAFEE modified for the K computer. This requires searching for candidate compounds that may work well among several millions of compounds while checking their chemical structures. Therefore, we collaborate with the Bio-IT Business Development Unit, Fujistu Limited which runs a computer-based drug design business, and selected several hundreds of candidate compounds for MP-CAFEE calculations. Candidate compounds include not only existing compounds. We also design novel compounds that work better. In such a case, we of course check whether it can be synthesized and is free of toxicity (side effects). It is an advantage of IT drug discovery that it can reduce development costs and time. However, its maximum benefit is that we can design various new compounds and predict candidate compounds through molecular simulation without being limited by existing compounds.

In this fiscal year, I believe we can produce a series of drug candidate compounds against relatively small target proteins. Because of computation time, we give priority to the surest thing, and then prepare to handle larger and more difficult targets in the future. We will practically complete this project in three years. The biggest goal of our project is to develop several compounds that may be studied in clinical trials by then. Of course it includes the results we will produce in this fiscal year.

As I mentioned before, it is one of the objectives of this project to get Japanese pharmaceutical companies interested in IT drug discovery that has been failing to produce successful results, and to have them work with us. Thanks to worldwide interest and the completion of the K computer, we have received favorable responses and offers from pharmaceutical companies. Since it is a national project, we have talked with many pharmaceutical companies about collaboration. Many companies have already given us various proposals. Some of them wish to start during this fiscal year. Since we are not yet ready to get off the block, we are now continuing discussions with them and are preparing for the future.

Lecture on Computational Life Sciences for New undergraduate Students

HPCI Program for Computational Life Sciences, RIKEN

Chisa Kamada

This February, Mr. Yukihiro Eguchi, Deputy-Program Director of Field One of the Strategic Programs for Innovative Research (hereafter "Field One"), delivered an invited lecture in the "4th Science Fair in Hyogo" at the Kobe International Exhibition Hall. Following the presentation, Professor Kuniyoshi Ebina asked that Field One provide a Special Lecture for the Kobe University subject, entitled "Invitation to Human Development Interdisciplinary Subjects". This compulsory course targets undergraduate students entering the Faculty of Human Development.

The Field One Special Lecture was delivered on June 29 and again on July 13 by Mr. Eguchi (293 attendees in total: 140 male, 153 female). The Kobe University Faculty of Human Development encompasses both the liberal arts and the sciences, and addresses various issues on human development and its surrounding environment. Students in the faculty range from those who major in astrophysics and computational biology to those who major in art, psychology, and sports science. They are also a diverse and representative group of Japanese young people in terms of intellectual curiosity. The title of the Field One Special Lecture was "What are Interdisciplinary Subjects?—Using Examples of the Integration of Biology with Physics, Chemistry and Mathematics through the Mediation of the K Computer". Mr. Eguchi's dual focus was the K computer and its versatility, and the importance of interdisciplinary science.

Field One is continuously engaged in educational and outreach activities for high school, college, and graduate school students. Most of the students we encounter major in one of the sciences and are already interested in computers. With the Field One Special Lecture, we were able to reach students with little or no previous exposure to the sciences and the K computer. As a result, we saw a big difference in the understanding of students before and after the lecture.

The most striking response from the students as to supercomputers and the K computer was that they believed supercomputers were different from what they had expected. More than ten percent of students thought that supercomputers had nothing to do with daily life. Those already aware of supercomputers had wondered what kind of change they could expect from improvements in computational speed, until they learned about the versatility of the K computer and its applicability to diverse simulation challenges. Some respondents described a negative image of supercomputers, because they often heard in the media that investment in supercomputers represents a waste of tax money. This Special Lecture offered a good opportunity for us to explain the importance of the K computer in society to such students.

The second most common reaction from students was that they were surprised to discover that supercomputers developed during the past twenty years are impacting our life today, and the prediction that the technology of today's supercomputers will become more common in daily life. Knowing the history and background of supercomputers, they were also surprised about the rapid progress of science and technology, and became more interested in supercomputers. One of the feedback comments, from which one can appreciate how impressions were changed by the lecture, is shown below:

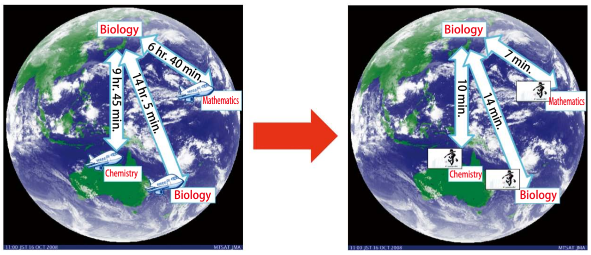
> Old supercomputers have been transformed into the personal computers and mobile phones that we use today, and the state-of-the-art K computer is destined to become a familiar part of our daily life within the next twenty years. In turn, new supercomputers will be born, and we will enter an unprecedented era in modern technology. Just the thought of it is exciting. –excerpted comment–

During the lecture, Deputy-Program Director Eguchi explained that biology is not a science separate from mathematics, physics, and chemistry, but complements them, and that the K computer can further strengthen such connections. To supplement the explanation, he provided pattern diagrams using global maps. Such notions seemed very new to many of the students, who had just passed university entrance examinations, and all appeared surprised. When the lecture concluded, one liberal arts student stated "It can be said that science supports progress in the world now, but I think it is important to integrate liberal arts into science. What's your opinion?" Mr. Eguchi responded that real-life problems could not be solved by science alone, and described the importance of interdisciplinary subjects, illustrating his answer with examples of approaches to create a safe and secure society. As previously described, the focus of the Faculty of Human Development is interdisciplinary. It was apparent that more than a few of these newly entered students felt anxiety about college life, because of the challenges posed by the more diversified curriculum than undergraduate students at other faculties of Kobe University. It was gratifying to see all participants better understand the importance of acquiring a broad undergraduate education.

Thanks to Professor Naoko Shirasugi of Kobe University, Field One received feedback about the lecture. It was a good opportunity for us to hear the students' direct and honest opinions. It appears likely that the understanding of the relationship between supercomputers and society shared by many students reflects a broader view within Japanese society. In order to win public understanding of the need for supercomputer technologies, it is important to make efforts to "familiarize" the public with supercomputers, even reaching those with no immediate interest in the subject. We hope the activities of Field One will help nurture young people, who will forge our future.

Finally, Field One wishes to extend its deepest appreciation to those involved from the Faculty of Human Development, Kobe University, including Professor Hounoki, Dean of the Faculty, Professors Ebina and Shirasugi, and, most of all, the 293 undergraduate students.

Biology is a complementary discipline to mathematics, physics and chemistry!



Pattern diagram: The K computer can strengthen ties between academic fields.

(http://www.data.kishou.go.jp/obs-env/portal/chishiki_ondanka/p01.html)

As part of efforts to strengthen educational and outreach activities, the lectures planned and given by the Faculty of Human Development, Kobe University and Field One in collaboration with RIKEN Advanced Institute for Computational Science and the other four strategic fields in the Strategic Programs for Innovative Research, are as follows:

"Special Lecture on Natural Environmental Science D" in the second semester of FY 2012 - New science opened up by the K computer-

Nov. 17 "Computing Machinery and the K computer in Society"
RIKEN Advanced Institute for Computational Science

"History and Changes of the Magnificent Natural Environment Surrounding Us"<Field 5>

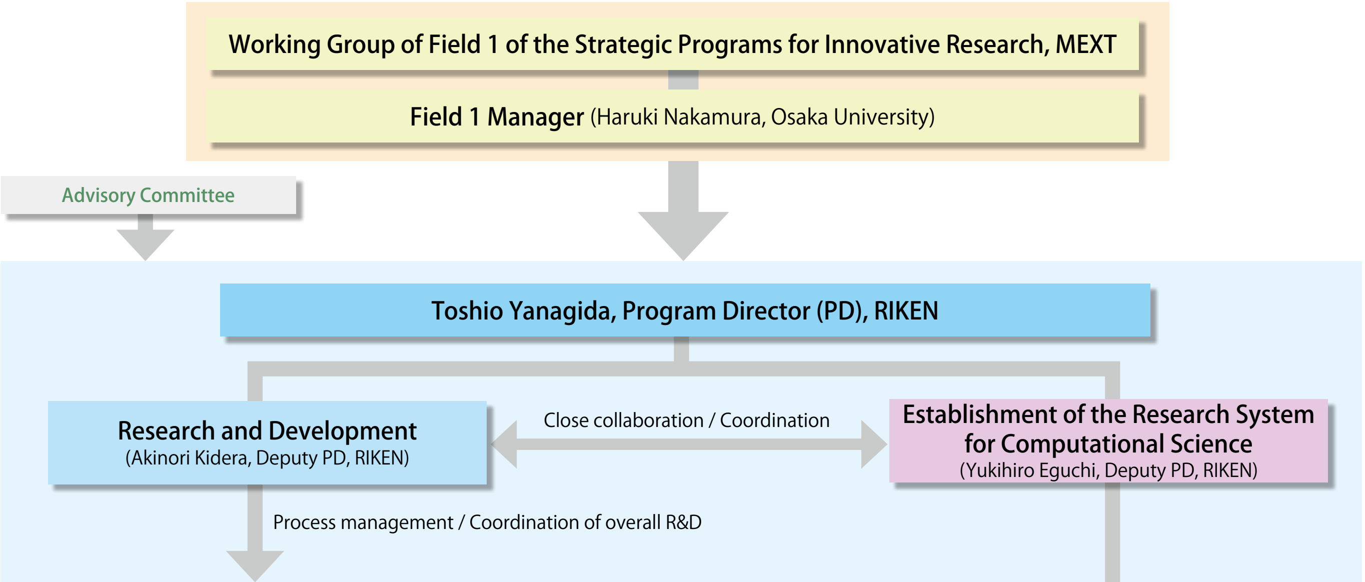
Dec. 1 "How to Assess the Severity of Energy Problems, and How to Overcome Them"<Field 2>

"The Mystery of Microorganisms Coexisting with Humans"<Field 1>

Dec. 15 "The Environment Created by Humans, its Safety, and its Soundness" <Field 4>

"Roles of the K computer in the Study of Massive Ocean-trench Earthquakes and Disaster Mitigation"<Field 3>

Research System for Computational Science



Working Group of Field 1 of the Strategic Programs for Innovative Research, MEXT

Field 1 Manager (Haruki Nakamura, Osaka University)

Advisory Committee

Toshio Yanagida, Program Director (PD), RIKEN

Research and Development
(Akinori Kidera, Deputy PD, RIKEN)

Close collaboration / Coordination

Establishment of the Research System for Computational Science
(Yukihiro Eguchi, Deputy PD, RIKEN)

Process management / Coordination of overall R&D

Theme 1: Simulations of biomolecules under cellular environments

- **Yuji Sugita**, GL, RIKEN Advanced Science Institute

Koichi Takahashi, RIKEN
Motonori Ota, Nagoya University
Ryuichiro Ishitani, The University of Tokyo
Hidetoshi Kono, Japan Atomic Energy Agency

Theme 2 : Simulation applicable to drug design

- **Hideaki Fujitani,** GL, Research Center for Advanced Science and Technology, The University of Tokyo

Noriaki Okimoto, RIKEN

Theme 3 : Hierarchical integrated simulation for predictive medicine

- **Shu Takagi,** GL, School of Engineering, The University of Tokyo

Yoshihiko Nakamura, The University of Tokyo
Kenji Doya, Okinawa Institute of Science and Technology Graduate University
Taishin Nomura, Osaka University

Theme 4 : Large-scale analysis of life data

- **Satoru Miyano,** GL, Institute of Medical Science, The University of Tokyo

Yutaka Akiyama, Tokyo Institute of Technology
Kiyoshi Asai, The National Institute of Advanced Industrial Science and Technology
Hideo Matsuda, Osaka University
Takashi Gojobori, National Institute of Genetics

High Performance Computing Development Group

- **Makoto Taiji,** GD, RIKEN

1) Management of computational resources
2) K computer-use support in Field 1

Planning and Coordination Group

- **Satoru Tomita,** GD, RIKEN

3) Human resource development
4) Establishment of human networks
5) Dissemination of research outcomes
6) Interdisciplinary projects
7) Holding of steering group meetings and holding of external advisory committees

Steering Committee

Coordination of project management through regular meetings

(PD) Program Director
(Deputy PD) Deputy Program Director
(GL) Group Leader
(GD) Group Director

Partnership

Use of HPCI computing resources

Outside organizations

Molecular biology, cell biology, biological physics, leading-edge experimental facilities, pharmaceutical companies, hospitals, various databases, etc.

Event Information

Announcement of the ISLiM International Symposium

4th Biosupercomputing Symposium

- Date : December 3(Mon) -5 (Wed), 2012
- Location : Tokyo International Forum (Chiyoda-ku, Tokyo)
- Conference fee : Free (get-together fee is charged.)

For more details, please see the website http://www.csrp.riken.jp/4thbscs/top_e. html

Academic-Industrial Collaboration in the Supercomputer "K Computer" and Drug Discovery/Medical Care

―HPCI Program for Computational Life Sciences―

- Conference fee: Free

For more details, please see the website http://hpci.me.es.osaka-u.ac.jp/

Third Seminar

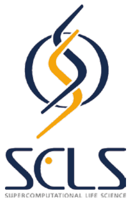
- Date: December 19 (Wed), 2012, 13:00 - 17:00
- Location: Umeda Center Building (Kita-ku, Osaka)

Fourth Seminar

- Date: January 25 (Fri), 2013, 13:00 - 17:00
- Location: Fukuracia Tokyo-Station (Chiyoda-ku, Tokyo)

News

The logo design for Field 1 was completed



SCLS
SUPERCOMPUTATIONAL LIFE SCIENCE

Logo design concept

The monogram combining the two letters of "S" in the motif of the double-helical structure of DNA symbolizes the life sciences, and a circle behind the monogram symbolizes the "cell = life" and "circle = harmony and coordination (cooperation)". Navy blue which is the color suggestive of intelligence represents supercomputers, and orange which is the color suggestive of upward mobility and liveliness represents life phenomena and life sciences. "SCLS" is an acronym for the theme of Field 1 "Supercomputational Life Science" and is pronounced as "esukurusu".
Please also see the website of Field 1 (http://www.kobe.riken.jp/stpr1-life/outreach/pr/logo_stpr1.html)

The Development and Use of the Next-Generation Supercomputer Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT)

Next-Generation Integrated Simulation of Living Matter

ISLiM
Integrated Simulation of Living Matter

The "Next-Generation Integrated Simulation of Living Matter" is a project sponsored by the Ministry of Education, Culture, Sports, Science and Technology(MEXT), in which research and development of simulation software to understand various phenomena occurring in the biological systems, including molecules and the human body, have been undertaken to realize a petascale simulation by making full use of the performance of supercomputer "K computer".

Strategic Programs for Innovative Research Field 1

Supercomputational Life Science

SCLS
SUPERCOMPUTATIONAL LIFE SCIENCE

SPIRE (Strategic Programs for Innovative Research) is a MEXT program aiming to produce groundbreaking results in computer science technology by maximizing the benefits of the HPCI (High Performance Computing Infrastructure) centered on the supercomputer "K computer", and encouraging developments in five research fields that need to be strategically addressed.
"Supercomputational Life Science" has conducted research with the mission of understanding and predicting life phenomena based on large-scale simulation and advanced data analysis, and to apply these results to design and implement medicine and medical care with its research.

BioSupercomputing Newsletter Vol.7 2012.12

Issued: December 2012

RIKEN

Research Cluster for Innovation
Computational Science Research Program
2-1 Hirosawa, Wako, Saitama 351-0198, Japan
Tel : +81-48-462-1488 Fax : +81-48-462-1220
http://www.csrp.riken.jp

HPCI Program for Computational Life Sciences
7-1-26 Minatojima-minamimachi, Chuo-ku, Kobe,
Hyogo 650-0047, Japan
Tel : +81-78-940-5692 Fax : +81-78-304-8785
http://www.kobe.riken.jp/stpr1-life