Eun-Sung Art Cho
Center for Biomolecular Simulations
Mail Code 3158
Columbia University
New York, NY 10027

January 23, 2004

Professor Rob de Ruyter van Steveninck
Biocomplexity Institute
Indiana University
Bloomington, IN 47405-7105

Dear Professor de Ruyter van Steveninck:

I hereby would like to apply for the recently advertised tenure-track Assistant Professor position in the Department of Physics at MIT. I am currently a postdoctoral research scientist in the department of chemistry and Center for Biomolecular Simulations at Columbia University. I have graduated with Ph.D. in physics from Brown University in June 2001. Before coming to Columbia in October 2002, I worked as a postdoctoral scholar in the Materials and Process Simulation Center at California Institute of Technology since July 2001 till October 2002.

My primary research interest lies in development of applications of multiple scale molecular docking. This interest incorporates quantum mechanics, molecular mechanics, Monte Carlo methods, statistical mechanics, and molecular dynamics all in one framework. I am mostly interested in bio-nanotechnological applications of this development such as motor proteins. I have been exposed to various fields of computational physics/chemistry throughout my graduate and postdoctoral work. They include Monte Carlo methods, wavelet formulation of path integral Monte Carlo, *ab initio* calculations of transition metals, carbon nanotube simulations, protein-ligand docking, and QM/MM (quantum mechanical/molecular mechanical) calculations. It is my intention to put all of these elements together to form one big framework, which will be applied to a lot of fascinating subjects such as bio-nanotechnology.

I am very much interested in teaching undergraduate students and leading them into research. I have had experiences of working with undergraduate students at Brown and Caltech and it was very encouraging for them and myself that we actually had real research output. The attached statement of teaching interests will give more details.

I look forward to hearing from you.

Sincerely,

Eun-Sung Art Cho

Attached: 1) Statement of research interests
2) Statement of teaching interests

Multiple Scale Molecular Docking in Functional Proteomics: A Tool for Bio-Nanotechnology

Eun-Sung Art Cho

Introduction

The theory and application of molecular docking has been the subject of intensive research over the last decade[1-3]. Docking is an important theoretical tool for proteomics. It addresses primarily the issue of functions of proteins in the genome. The first step in predicting the function of proteins is to predict where and how strongly various ligands bind to the protein. For this, there have been a handful of academic and/or commercial programs developed over the years[4-9]. Although there are certain standards currently employed, which pharmaceutical industry had some success in applying to drug discovery effort, the research in this field is quite open and active. Moreover, the application of docking is not limited to drug discovery. For example, it is suggested that docking be used as a tool for studying protein-folding problem[10].

Just like many chemical phenomena, docking involves multiple levels of scales in both time and length (Figure 1). Different physical/chemical properties can be related to different length scales. At the smallest scale, quantum mechanics dominates and electronic structure is what one is describing with simulations. In docking, this can be translated into polarization of atoms of the binder in binding sites. On the larger scale, the conformation of the receptor, which is usually a large protein, can be described by force-field calculations using Monte Carlo (MC) or molecular dynamics (MD). At the even larger scale, one can consider an ensemble of molecule complexes, which involves statistical mechanics. This last part is in a sense the ultimate way to describe experimental reality.

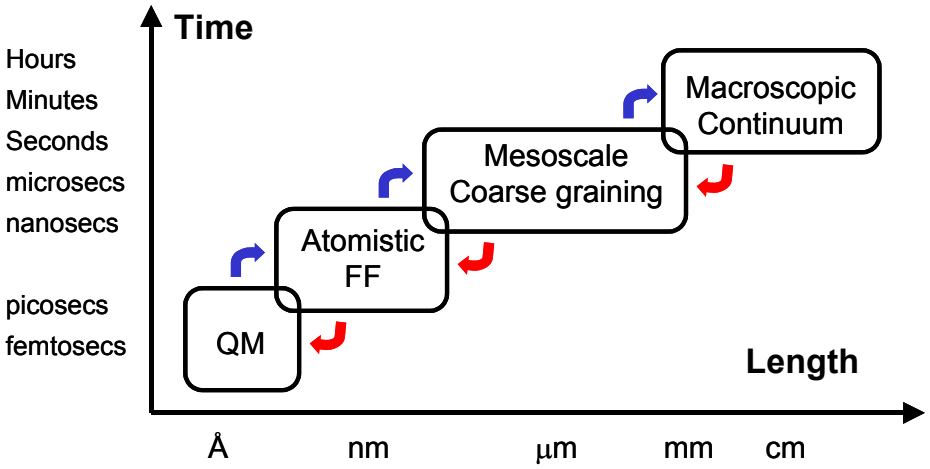

Figure 1. Hierarchy in computational modeling

It is quite possible to make application of current stage of docking methods to practical research problems. However, the problem of molecular docking is far from being solved. It is certainly true that, especially with the growing computing power, there is

much room to improve in methodology of docking. Therefore, my proposed research is two-fold: docking methodology and application of docking.

1. Docking Methods

1.1 Improvement in scoring

Recently, I have investigated adopting QM/MM (quantum mechanics/molecular mechanics) methods to docking problems[11]. The aim of this investigation was to evaluate the importance of electric charges in docking, which is an issue that has been overlooked. In many of the cases tried, I found that accurate charge (variable) values could be important in obtaining the right configurations and binding energies. With the procedure I have developed, I plan to complete a docking program which can be used easily by researchers in other fields. Further refinement of this method is to freely control the region where quantum mechanics should be imposed. The current setting I have is to confine the QM region to be ligand alone. However, this idea can be extended to include parts of the protein within the binding sites, or break up the binder and include only parts of it in case it is large (e.g. enzyme). The immediate problem that arises when this is done is how to calculate the change in charges. For this, I propose developing a new quantum mechanical charge calculation algorithm based on ESP (electrostatic potential) method incorporating extrapolations.

This program development will be done either in conjunction with existing docking programs or as a stand-alone program. This part of my proposal is about improving scoring in docking and alludes to the smallest scale of docking.

1.2 Search algorithm

Search algorithm in docking problem is an important issue and has been exploited quite extensively[12,13]. I have developed and implemented the concept of completeness with clustering and enrichment to guarantee complete search of the conformational space one is interested in[14]. I propose to further enhance this development and combine it with MC and MD to make the conformational search more complete. This part is in the scale of atomistic force-field calculations.

All of these algorithms for search and scoring are based on the assumption that the energy landscape near the native (experimental) structure is thoroughly sampled and the native structure gives the global free energy minimum. However, in reality, molecular complexes are thermodynamic ensembles of structures. This brings in statistical elements to docking and I propose to develop an algorithm in which one generates an ensemble of structures using MD at various temperatures and link them to find a statistical approximation to the experimentally observed structures. In case of HIV-1 protease and DHFR (dihydrofolate reductase), a rather large number of crystal structures have been determined and these data can be used to form conformer ensembles[15-20]. I plan to utilize these data to guide in developing the statistical method of docking.

1.3 Polarizable force-field docking

QM/MM calculations can be costly, and in an industrial setting might not be appropriate for procedures such as VLS (virtual ligand screening), where speed is important in order to scan through a large database. A cheaper alternative, which will retain the

benefit of variable charges, would be using polarizable force field. I propose to develop a docking method which involves polarizable force field and evaluate the effects of it. The result with QM/MM calculations indicates that the use of polarizable force field in docking will bring some benefits in finding the right configurations and binding energy. A positive result will extend this project to developing a docking program which utilizes polarizable force field.

2. Applications of Docking

1.1 Enzyme design

The difference in deacylation rates for acyl-enzyme intermediates in penicillin-binding proteins (PBP) and β-lactamases is a key phenomenon in understanding bacterial antibiotic resistance. There has been a computational work on this subject at Columbia using QM/MM method[21] without proper docking procedure. Using B3LYP DFT (density functional theory) level of quantum mechanics, the authors tried to elucidate the detailed energy pathways of these systems. My QM/MM docking method should be able to incorporate proper docking into this problem. It is my plan to modify my method so that transition states can be calculated in order to integrate all of the theoretical considerations for this problem and other enzyme design problems.

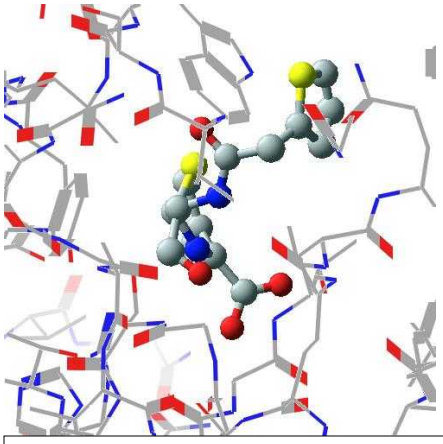

Figure 2. R61 (PBP) enzyme with cephalothin ligand in the binding site

1.2 Molecular motor

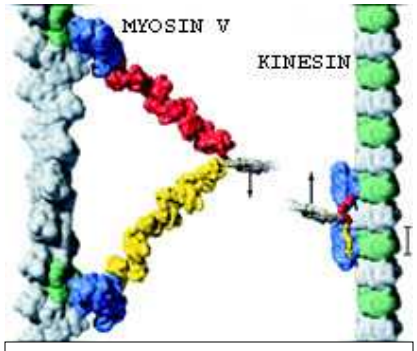

Figure 3. Myosin V and kinesin molecular motors (Vale, UCSF)

Molecular motors are building blocks of life and have been a subject of research for a very long time[22,23]. Only recently, scientists are gaining understanding and prospect of application in this nature's nanomachine. Myosin V and kinesin are two primary examples of molecular motors. These tiny motors move along filamentous tracks that are ubiquitous within every living cell. Although these processes are different from traditional docking in that 1) it is a dynamical process and 2) there is no stationary binding site, they are the same as docking in that they involve two separate molecules that are weakly bound. I propose to use a docking program as a basis upon which one can simulate a dynamical process, in particular, molecular motor motion in this case, by adding interaction terms. The benefit of doing this is the conceptual ease of designing nanomachines, which follow the same principle as molecular motors.

Miscellaneous Projects

3.1 Parallel programming on Linux cluster systems

It is increasingly becoming clear that the most economical way to acquire high computing power in an academic setting is to build a Linux cluster system. I have an extensive experience with parallel computing in general and worked on building Beowulf clusters at Samsung Advanced Institute of Technology. I programmed extensively with PVM (parallel virtual machine) and MPI (message passing interface). I would like to utilize my expertise in parallel computing to maximize the university resources for computing power that is essential to today's scientific research. This project will include building a customized Linux cluster system and parallelization of both existing codes and codes that I will be developing.

3.2 Binding sites and crystallographic water

There are some protein-ligand structures in PDB (protein data bank) that have common protein as receptor. One example is the family of trypsin co-crystals, for which there are currently eight of them in PDB. Careful examination of these structures reveals that some of the crystallographic waters are common in all of them. This fact suggests that there may exist some inherent structural properties, which make the water molecules to remain in the binding sites. Being able to predict the location of the crystallographic waters will greatly help docking problem. There has been an attempt to devise a theoretical tool to identify these "buried waters"[24]. However, it seems that extensive MD simulations would be required to do a rigorous approach to finding these stable waters. This amounts to using rather large amount of computing time. It will involve simulating explicit water molecules along with the protein atoms. By defining the "stability condition" of waters throughout the MD simulation one should be able to identify those waters that inherently remain in the binding pockets during docking. Linux clusters and parallel programs for MD can provide one with an inexpensive way to perform such simulations.

3.3 Homology modeling of transmembrane proteins

Transmembrane proteins are peculiar species in that they either are attached to or traverse the lipid bilayer, which is a hydrophobic environment. They mediate our senses, and are involved in cell recognition and communication processes and they have emerged as a prominent superfamily for drug targets[25,26]. However, it is extremely difficult to crystallize these proteins and only a handful of structures are known. Therefore, it is essential to utilize homology modeling to model this class of proteins. At Columbia University, there has been a development effort for a homology modeling code called PLOP (protein local optimization program)[27]. I propose to use this program as a platform and incorporate new scoring functions, which are suitable for hydrophobic region. Eventually, this project will be combined with molecular docking development to give a theoretical prediction of transmembrane protein functions.

3.4 Application of wavelet formalism of path integral Monte Carlo

Wavelet is a great tool for multi-scale analysis. For that reason, it is generally believed that wavelet can bring fundamental changes in computational physics/chemistry. Although this belief has not yet become reality, I still maintain my position that wavelet is

very useful in formulation of a lot of important concepts in simulations of atoms and molecules. I have formulated "wavelet path integral Monte Carlo"[28] during my Ph.D. study at Brown. I plan to return to this theme and extend the experiment to more realistic systems. I would also like to develop the so-called "wavelet Ewald summation". The idea is to find a right kind of wavelets to use instead of Gaussian functions for Ewald summation. Once successfully implemented, the advantage of this formulation is the great reduction in size of basis set needed and computing time.

References

(1) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Current Opinion in Chemical Biology* **2001**, *5*, 375-382.

(2) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078-1082.

(3) Lengauer, T.; Rarey, M. Computational methods for biomolecular docking. *Current Opinion in Structural Biology* **1996**, *6*, 402-406.

(4) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-Based Molecular Design. *Accounts of Chemical Research* **1994**, *27*, 117-123.

(5) Friesner, R.; Halgren, T. A.; Shenkin, P.; Murphy, R. B.; Banks, J. et al. Flexible ligand docking with glide: Evaluation of structure prediction performance and utility for virtual ligand screening. *Abstracts of Papers of the American Chemical Society* **2002**, *223*, U465-U465.

(6) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **2001**, *308*, 377-395.

(7) Kramer, B.; Metz, G.; Rarey, M.; Lengauer, T. Ligand docking and screening with FlexX. *Medicinal Chemistry Research* **1999**, *9*, 463-478.

(8) Abagyan, R.; Totrov, M.; Kuznetsov, D. Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry* **1994**, *15*, 488-506.

(9) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins-Structure Function and Genetics* **2003**, *52*, 609-623.

(10) Crick, F. H. C. The packing of α-helices: simple coiled-coils. *Acta Cryst* **1953**, *6*, 689-697.

(11) Cho, A. E.; Guallar, V.; Berne, B.; Friesner, R. The importance of electric charges in molecular docking - QM/MM approach. *In preparation* **2004**.

(12) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S. et al. Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 731-751.

(13) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks, C. L. Assessing search strategies for flexible docking. *Journal of Computational Chemistry* **1998**, *19*, 1623-1631.

(14) Cho, A. E.; Wendel, J. A.; Vaidehi, N.; Kekenes-Huskey, P. M.; Floriano, W. B. et al. The MPSim-Dock Hierarchical Docking algorithm: Application to the eight Trypsin Inhibitor co-crystals. *Submitted to Journal of Computational Chemistry* **2004**.

(15) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D. et al. Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *European Journal of Biochemistry* **1977**, *80*, 319-324.

(16) Vondrasek, J.; Wlodawer, A. Database of HIV proteinase structures. *Trends in Biochemical Sciences* **1997**, *22*, 183-183.

(17) Vondrasek, J.; vanBuskirk, C. P.; Wlodawer, A. Data base of three-dimensional structures of HIV proteinases. *Nature Structural Biology* **1997**, *4*, 8-8.

(18) Feeney, J. NMR studies of ligand binding to dihydrofolate reductase. *Angewandte Chemie-International Edition* **2000**, *39*, 290-312.

(19) Dolle, R. E.; Nelson, K. H. Comprehensive survey of combinatorial library synthesis: 1998. *Journal of Combinatorial Chemistry* **1999**, *1*, 235-282.

(20) Yamazaki, T.; Hinck, A. P.; Wang, Y. X.; Nicholson, L. K.; Torchia, D. A. et al. Three-dimensional solution structure of the HIV-1 protease complexed with DMP323, a novel cyclic urea-type inhibitor, determined by nuclear magnetic resonance spectroscopy. *Protein Science* **1996**, *5*, 495-506.

(21) Gherman, B. F.; Goldberg, S. D.; Cornish, V. W.; Friesner, R. Mixed Quantum Mechanical/Molecular Mechanical (QM/MM) Study of the Deacylation Reaction in a Penicillin Binding Protein (PBP) Versus in a Class C β–Lactamase. *Submitted to Journal of American Chemical Society* **2003**.

(22) Schliwa, M.; Woehlke, G. Molecular motors. *Nature* **2003**, *422*, 759-765.

(23) Schliwa, M. Protein transport: Molecular motors join forces. *Nature* **1999**, *397*, 204-205.

(24) Rashin, A. A.; Iofin, M.; Honig, B. Internal Cavities and Buried Waters in Globular-Proteins. *Biochemistry* **1986**, *25*, 3619-3625.
(25) Bergsma, D. J.; Wilson, S. G-protein-coupled receptors. *Biofutur* **1999**, 26-27.
(26) Dong, X. Z.; Han, S. K.; Zylka, M. J.; Simon, M. I.; Anderson, D. J. A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. *Cell* **2001**, *106*, 619-632.
(27) Jacobson, M.; Friesner, R. PLOP (Protein Local Optimization Program): A software platform for physics-based, all-atom predictive protein modeling. *In preparation* **2003**.
(28) Cho, A. E.; Doll, J. D.; Freeman, D. L. Wavelet formulation of path integral Monte Carlo. *Journal of Chemical Physics* **2002**, *117*, 5971-5977.

Statement of Teaching Interests

Eun-Sung Art Cho

Although in today's universities research has become more important than teaching, it is my belief that the true value of a university lies in education of the members of the society within which it exists. One can imagine that research can be done more efficiently at places like government labs and industrial research centers where all the activities are geared towards producing results. The mission of universities therefore must be something more than research. I personally take great pleasure in passing my knowledge onto other people, no matter in what form: teaching a lab section or helping a friend solving math problems in a dormitory. This has been so since I was as young as in elementary school. It is why I am applying for a teaching position.

It goes without saying that one should teach with his/her forte. Having been educated in physics and mathematics, my strength in teaching lies in subjects for which mathematical or physical reasoning is important. I am able and willing to teach courses in general physics, general chemistry, mathematical methods for physical sciences, and computational sciences. More advanced courses could include Monte Carlo methods in physical sciences, statistical mechanics, and computational methods in physical sciences.

I also propose to develop an exciting new course, which will involve some of the brightest undergraduate students as well as graduate students. This course will be structured under the general framework of computational methods with lab work and students will be doing more computer lab work than reading textbooks and solving homework problems. Although there will be some lectures covering various computational methods for simulations, students will be completely free on the choice of computational methods. They will have opportunities through which they can taste the true feeling of cutting-edge research. Each term, the topics will be selected according to both the interests of students and the trend of current research. Rarely would one find the same topics in two different terms. This plan stems from my experiences with undergraduate students at Brown and Caltech. As a Ph.D. student and a postdoc at these places, I had the opportunity to work with undergraduate students. At Brown, I helped a student to perform first principle calculations for his experiment; at Caltech, I guided a student to do some supplementary calculations for me. Although their level of knowledge in physics and math was not at that of an advanced graduate student, being a young generation growing up with computers from a very early age, they were able to plow through their tasks. I can honestly say I was impressed and we all benefited from the collaborations. The reasoning behind my proposal is that instead of letting students learn computing for the need of doing physical sciences, one can interest students who already have computer skills into using the skills to do research in other sciences. Depending on how the students react, I imagine that some of the results can be made into publications and that can be a real exciting motivation for students. My philosophy about doing science is that one should do it because one finds joy in it. And I believe the best way to let the students see a glimpse of such joy is to help them produce tangible results, no matter how small they could be.