Final Report Certification
for
CRADA Number IBM 0617

Between

UT-Battelle, LLC

and

IBM
(Participant)

Instructions:

Mark the appropriate statement in 1a or 1b below with an 'IX." Refer to the articles in the CRADA terms and conditions governing the identification and marking of Protected CRADA Information (PCI).

If no PCI is identified, the report will be distributed without restriction. If PCI is identified, the report distribution will be limited in accordance with the CRADA terms and conditions governing release of data. In all cases items 2 and 3 must be true. That is, the report cannot contain Proprietary Information and a disclosure must be filed prior to release of the report.

This certification may either be made by using this form or may be made on company letterhead if the Participant desires. A faxed copy of this completed form is acceptable.

The following certification is made for the subject final report:

1. (a) □ The final report contains information that qualifies as "Protected CRADA Information" (PCI). The PCI legend is printed on the report cover, and the PCI is clearly identified.

OR

(b) ☒ The final report does not contain "Protected CRADA Information." The "Approved for Public Release" legend is printed on the report cover.

2. The final report does not contain Proprietary Information.

3. By the signature below, the Participant has no objection to the public distribution of the final report due to patentable information.

For the Participant:

LOU GUTTMAN
(Name)

EXECUTIVE, GOVERNMENT PROGRAMS
(Title)

6/27/07
(Date)

From: Scott, Robert E.
Sent: Wednesday, May 23, 2007 7:08 AM
To: McCoy, Debbie D.; Porto, Casey
Subject: RE: IBM CRADA 617 final report

The final report certification requires an actual signature by the participant.

From: McCoy, Debbie D.
Sent: Tuesday, May 22, 2007 6:24 PM
To: Porto, Casey
Cc: Scott, Robert E.
Subject: IBM CRADA 617 final report

Dear Casey:

As requested, attached is the report on the IBM CRADA 617. The pdf cover file seems to be corrupted such that it will not go into a single document; hence, the cover in one file and the final report in another.

Best regards,
Debbie

****
Debbie McCoy
Computing and Computational Sciences Directorate
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6163
Voice: 865 574-6185
Fax: 865 574-4839
Bldg. 5700, MS 6163
mccoydd@ornl.gov
http://computing.ornl.gov

Final Report Certification
for
CRADA Number IBM 0617

Between

UT-Battelle, LLC

and

IBM
(Participant)

Instructions:

Mark the appropriate statement in 1a or 1b below with an 'IX." Refer to the articles in the CRADA terms and conditions governing the identification and marking of Protected CRADA Information (PCI).

If no PCI is identified, the report will be distributed without restriction. If PCI is identified, the report distribution will be limited in accordance with the CRADA terms and conditions governing release of data. In all cases items 2 and 3 must be true. That is, the report cannot contain Proprietary Information and a disclosure must be filed prior to release of the report.

This certification may either be made by using this form or may be made on company letterhead if the Participant desires. A faxed copy of this completed form is acceptable.

The following certification is made for the subject final report:

1. (a) □ The final report contains information that qualifies as "Protected CRADA Information" (PCI). The PCI legend is printed on the report cover, and the PCI is clearly identified.

OR

(b) ☒ The final report does not contain "Protected CRADA Information." The "Approved for Public Release" legend is printed on the report cover.

2. The final report does not contain Proprietary Information.

3. By the signature below, the Participant has no objection to the public distribution of the final report due to patentable information.

For the Participant:

Al Geist
(Name)

Corporate Fellow
(Title)

09/15/2006
(Date)

Final Report for CRADA # 0617.02 with IBM

Computational Biology, Advanced Scientific Computing, and Emerging Computational Architectures

Summary:

This CRADA was established at the start of FY02 with $200K from IBM and matching funds from DOE to support post-doctoral fellows in collaborative research between International Business Machines and Oak Ridge National Laboratory to explore effective use of emerging petascale computational architectures for the solution of computational biology problems. "No cost" extensions of the CRADA were negotiated with IBM for FY03 and FY04. The CRADA expired in October 2004. In FY02 we hired two post-docs, Stephen Shevlin and Pratul Agarwal. Stephen Shevlin came directly from IBM's Zurich Research Laboratory where he was a post-doc in the Computational Biochemistry and Material Science group doing electronic structure computations. Pratul Agarwal came with strong recommendations from Pennsylvania State University. His expertise is computational biology as applied to molecular dynamics. Working with IBM researchers, he has been successful at scaling AMBER on the IBM Blue Gene supercomputer. In FY04 Stephen Shevlin was hired as an ORNL staff member and Tema Fridman replaced him as a post-doc on this CRADA. In her short tenure, she developed a new method for searching protein sequence databases for peptide identification that was two orders of magnitude faster than the traditionally used SEQUEST program.

The next three sections describe the research activities of these three post-doctoral fellows while they worked under this CRADA. This is followed by a list of publications generated as a result of this research.

Stephen Shevlin

Stephen Shevlin worked on the capacitive properties of monolayer-protected clusters. Nanometer-sized particles are interesting because of radically different catalytic, electronic and optical properties, when compared to the bulk. This is because the crystal structure of these nanoparticles is often different from that of the bulk material, with commensurate differences in the electronic structure (and related optical and chemical properties). It is important to elucidate the fundamental interplay between atomic structure and electronic structure in order to gain a deeper understanding of these properties, with a view towards designing nanoscale devices that can function as optoelectronics or medical biosensors.

An important prototypical system for the study of these nanoparticles are the monolayer protected metal clusters (MPCs), because they are easy to fabricate, do not aggregate, and are stable. In particular, gold MPCs passivated by a monolayer of thiol molecules have been investigated experimentally by electrical, optical, and for the larger diameter cores, structural means. However, the understanding of these properties, especially for the smaller clusters, is lacking.

Dr. Shevlin studied a typically interesting gold cluster size, the 38 atom gold cluster (Au_{38}), both bare and passivated ($Au_{38}[SCH_3]_{24}$), including their structural motifs, passivation effects, and the charging characteristics. To accurately capture the quantum mechanical nature of these gold clusters, calculations were performed on the IBM-SP3 "Eagle" and IBM-SP4 "Cheetah" mainframe computers based here at ORNL, using the Carr-Parrinello-Molecular-Dynamics density functional methods to calculate both the atomic and electronic structure.

Structurally he found that the global minima for both the bare and passivated clusters, is a "disordered" tetrahedral structure, rather than the expected truncated octahedron structure. He found that the effects of surface passivation decrease this barrier to cluster distortion by increasing the surface stress, passivated clusters are much more "disordered" than bare clusters. He also found that the charging characteristics of these clusters are strongly determined by the gold cores, and that the passivation layer acts as an additional dielectric that gives a larger capacitance. He learned that the energetics of the clusters can be treated as per the simple jellium model, and that the lowest energy minimum occurs when the cluster is charged with two excess electrons.

Pratul Agarwal

Pratul Agarwal did research in three complementary areas: understanding protein structure, function and dynamics through computational biology, computational modeling of Protein-DNA complexes, and high performance algorithms for protein conformation search.

Multi-scale modeling of enzyme cyclophilin A was performed to understand the role of protein dynamics in enzymatic catalysis. This enzyme is required for various cellular process including protein folding, signaling and transport. The role of protein dynamics of various conserved residues in the cis/trans isomerization catalyzed by cyclophilin A was investigated at femtosecond (10^{-15}) to millisecond (10^{-3}) time scales. These investigations provided new insights into protein structure, function and dynamics, which are not available from biochemical experiments. A network of protein vibrations promoting catalysis in cyclophilin A was identified and characterized. This network is composed of conserved residues, which are conserved across several species. There is medical interest in this enzyme, as its incorporation into HIV (type-1) is required for the infectious activity of the virus. The protein-protein interactions between cyclophilin A and HIV-1 capsid protein have been investigated in detail, which provided valuable insights. The results of Dr. Agarwal's investigations were presented at The Biophysical Society Conference, March 2nd-5th 2003, San Antonio (TX) and published in the peer-reviewed journal *Proteins: Structure, Function, and Bioinformatics*. Another paper has also been submitted for publication in the *Journal of the American Chemical Society.*

Molecular modeling of protein-DNA complexes was performed to understand the nature of interactions within important complexes. The focus of Dr. Agarwal's projects was to investigate the role of dynamics and structure in the mechanism of biomolecular recognition used by DNA binding proteins. The results could enable prediction of DNA-binding sites for protein whose structure and target sequences are unknown. The following protein-DNA complexes were investigated:

I. Restriction endonuclease M. *Hha*I was investigated based on atomistic modeling of the protein-DNA complex, for preferential recognition of methylated cytosine in the target DNA sequence. Dynamic cross-correlation coefficients and protein-DNA interaction energy (a sum of electrostatic and van der Waals energy terms) were used to perform comparative investigations of bound DNA with non-methylated and methylated cytosine.

II. Transcription Factor SP1, whose 3-dimensional structure is unknown, has been investigated for prediction of binding sites. Based on homology modeling, an atomistic model has been prepared and dynamic cross-correlation and protein-DNA interaction energy calculations are being used to tabulate a list of binding sequences.

Manuscripts describing these investigations were submitted to peer-reviewed journals for publication as a full article.

Pratul Agarwal developed a new algorithm for efficiently searching the conformation space of a protein for local and global minima. His method is both high-performance and high-throughput. Combining parallel programming strategies and IBM's Blue Gene cellular architecture, this method searches the minimum energy conformation of a protein very rapidly. In addition, Pratul scaled AMBER (bio-molecular simulation package) on the IBM Blue Gene supercomputer at Argonne National Laboratory and identified areas of code for further optimizations. This work will enable long molecular dynamics simulations of complex biological systems. A manuscript was submitted to the *Journal of Parallel and Distributed Computing*, but it was not accepted for publication.

Tema Fridman

While working on the CRADA, Tema Fridman continued development of new computational algorithms for mass spectrometry analysis, focusing on tandem mass spectra identification.

SEQUEST represents one of the most widely used and accurate programs for peptide identification via database searches. Developed several years ago, SEQUEST was not specifically designed for large-scale applications. One of the limiting factors in meeting the needs for genome-scale applications is its (lack of) computing speed.

Under directorship of Ying Xu, with collaborators Jane Razumovskaya, Nathan Verberkmoes, and Greg Hurst, Tema Fridman developed a new method with high discriminating power for searching protein sequence databases for peptide identification. The accuracy compares favorably to the SEQUEST scoring function, with better separation between correct and incorrect matches. The algorithm was also found to be two orders of magnitude faster than the SEQUEST program.

The algorithm has been tested on a data set of 3771 experimental spectra that resulted from performing an LC-MS/MS experiment on a protein mixture of eight purified proteins. Our peptide database consisted of all possible tryptic peptides generated from the eight target proteins (having 803 peptides), and from the 2873 S. oneidensis proteins (having 289,166 peptides). The latter was used as a distractor dataset.

Among 1053 spectra of parent charge one, Luck algorithm identified 526 correctly versus 532 found by Sequest. Among the remaining 2721 spectra of parent charge two and three, we

identified 496 versus SEQUEST's 505 for parent charge two, and 221 versus SEQUEST's 227 for parent charge three. The spectra, that SEQUEST identified and Luck function did not, have very low Xcorr score. In terms of sensitivity -- specificity analysis, Luck function performs better than SEQUEST due to greater separation between correct and incorrect identifications.
Dr. Fridman and her collaborators published several papers on this work. These and the 10 other publications resulting from the research under this CRADA are listed below.

Journal Publications:

S. Shevlin, et al; "Covalent Attachment of Gold Nanoparticles to DNA Templates," *J. Nanosci Nanotechnology,* 2, pp. 397, 2002.

J. B. Watney, **P. K. Agarwal**, and S. Hammes-Schiffer, "Effect of Mutation on Enzyme Motion in Dihydrofolate Reductase," *J. Am. Chem. Soc.,* 125, pp. 3745-3750, 2003.

P. K. Agarwal, "Computational Studies of the Mechanism of Cis/Trans Isomerization in HIV-1 Catalyzed byt Cyclophilin A," *Proteins: Struct. Funct. Bioinformatics,* 56, pp. 449-463, 2004.

P. K. Agarwal, A. Geist, and A. Gorin, "Protein Dynamics and Enzymatic Catalysis: Investigating the Peptidyl-Prolyl Cis/Trans Isomerization Activity of Cyclophilin A," *Biochemistry,* 43, pp. 10605-10618, 2004.

A. Gorin, R. Day, A. Borziak, M. Strader, G. Hurst, and **T. Fridman**, "Probability Profile Method-New Approach to Data Analysis in Tandem Mass Spectrometry," **CSB2004** Conference Proceedings, pp.499-502, 2004.

T. Fridman, J. Razumovskaya, N. Verberkmoes, G. Hurst, and Y. Xu, "An Alternative to SEQUEST Cross-Correlation Scoring Algorithm for Tandem Mass Spectra Identification Through Database Lookup: The Luck Scoring Function, and the Probability of an Unrelated Spectra Match Model," **Currents in Computational Molecular Biology 2004, RECOMB**, p.66, San Diego, CA, 2004.

T. Fridman, J. Razumovskaya, N. Verberkmoes, G. Hurst, V. Protopopescu, and Y. Xu, "The Probability Distribution for a Random Match Between an Experimental-Theoretical Spectral Pair in Tandem Mass Spectrometry," *JBCB 2005*, 3, No. 2, pp. 1-22, 2005.

T. Fridman, V. Protopopescu, G. Hurst, A. Borziak, and A. Gorin,, "Generating Theoretical Spectra for Peptide Identification," **ETMBS '05** Conference Proceedings, p.180-189, Las Vegas, NV, June 20-23, 2005.

P. K. Agarwal, "Role of Protein Dynamics in Reaction Rate Enhancement by Enzymes," *J. Am. Chem. Soc.,* 127, pp.15248-15246, 2005.

P. K. Agarwal, "Enzymes: An Integrated View of Structure, Dynamics and Function," Invited Review Article: **Microbial Cell Factories**, 5, p. 2, 2006.

S. R. Alam, **P. K. Agarwal**, J. S. Vetter, and A. Geist, "Performance Characterization of Bio-Molecular Simulations Using Molecular Dynamics," **PPoPP Proceedings**, 2006, Accepted.