LA-UR-96- 2819

Title: **Computational and Theoretical Aspects of Biomolecular Structure and Dynamics**

Author(s): **Angel Garcia, T-10**
Joel Berendzen, P-21
Paolo Catasti, LS-8
Xian Chen, LS-8
Goutam Gupta, T-10
Gerhard Hummer,, T-10
Edward Kober, T-14
Tudor Opera, University of Timisoara, Hungary
Lawrence Pratt, T-12
Benno Schoenborn, LS-DO
Chang-Shun Tung, T-10
Santhana Mariappan

Submitted to: **DOE Office of Scientific and Technical Information (OSTI)**

MASTER

Los Alamos
NATIONAL LABORATORY

Form No. 836 R5
ST 2629 10/91

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DISCLAIMER

Computational and Theoretical Aspects of Biomolecular Structure and Dynamics

Angel E. Garcia*, Joel Berendzen, Paolo Catasti, Xian Chen, Goutam Gupta, Gerhard Hummer, Edward Kober, Tudor I. Opera (University of Timisoara, Hungary), Lawrence Pratt, Benno P. Schoenborn, Chang-Shun Tung, and Santhana Mariappan

Abstract

This is the final report of a three-year, Laboratory-Directed Research and Development (LDRD) project at the Los Alamos National Laboratory (LANL). This project sought to evaluate and develop theoretical, and computational bases for designing, performing and analyzing experimental studies in structural biology. Simulations of large biomolecular systems in solution, hydrophobic interactions, and quantum chemical calculations for large systems have been performed. We have developed a code that implements the Fast Multipole Algorithm (FMA) that scales linearly in the number of particles simulated in a large system. This code has been ported to CM-5 parallel computers and Crays. New methods have been developed for the analysis of multi-dimensional NMR data in order to obtain high resolution (atomic) structures. These methods have been applied to the study of repetitive DNA sequences found in the human centromere, and sequences linked to genetic diseases (Fragile X Syndrome). Combined theoretical, two-dimensional NMR, and calorimetric studies have been performed for centromeric sequences in order to determine the role of unusual base pairings in determining the stability of the secondary structure of centromeres. The structure and dynamics of myoglobin have been studied by molecular dynamics simulations to further analyze neutron diffraction data. Molecular dynamics simulations have given insight regarding the structure and dynamics of surface-bound water molecules. We have analyzed and implemented a continuum dielectric model for solvation of electrostatic interactions in aqueous solutions. The method developed to solve the requisite macroscopic Poisson equation permits explicit demonstration of numerical convergence and systematic utilization of coarse-grained results.

---

* Principal investigator, e-mail: angel@agua.lanl.gov

1. Background and Research Objectives

Major problems in structural biology are concerned with protein folding and with the structure, dynamics and function of proteins and protein assemblies. A detailed understanding of the mechanisms that determine the physico-chemical properties of biomolecules will ultimately reveal their structure and function. The understanding of the structure and function of biomolecules and biomolecular assemblies will be crucial for problems in biotechnology, such as the development of new drugs in the treatment of disease. Biomolecules and biomolecular assemblies are studied, at an atomic level, by means of multidimensional nuclear magnetic resonance (NMR) spectroscopy, X- ray and neutron diffraction, small angle scattering, and infrared (IR) and Raman spectroscopy. Biomolecular systems are extremely complex; and therefore, the analysis and interpretation of most experimental data requires performing computer simulations. These simulations require the most modern computers and advanced theoretical techniques.

The number of biological systems is extremely large (much greater than 10^{12}). Consequently, the proper choice of the system to be studied is extremely important. A wrong choice of system can make work extremely difficult and can make comparison with the results from other groups impossible, thereby limiting the impact of the work. An ideal system on which we would concentrate our theoretical and experimental efforts is myoglobin (Mb). It has been widely studied and has characteristics common to a great variety of proteins of various degrees of complexity. The structure, dynamics and stability of Mb are currently being studied by many groups at Los Alamos. This vast amount of information available to us will give a better understanding of the factors determining protein folding and dynamics. Experimental data will be employed to calibrate energy functions used in computer simulation studies. New experiments may be suggested and new theories and hypothesis can be tested in the computer.

2. Importance to LANL's Science and Technology Base and National R&D Needs

Simulations on biological macromolecules using molecular dynamics and Monte Carlo methods provide an important tool for basic research and biotechnology. Accurate and realistic simulations of proteins and other biomolecules in solution allow one to analyze the structure, dynamics, and function of a targeted biomolecule and predict such things as docking sites, interaction energies, and the effects of site-specific mutagenesis. The focus on Mb forms a link that encompasses computational, diffraction, laser spectroscopy, protein engineering and multi-dimensional NMR capabilities at Los Alamos. This work complements ongoing LANL

experimental and theoretical programs in protein and nucleic acid dynamics. It also enhances the visibility and reputation of the Laboratory in high performance computing, genomics, structural biology, and computational biology.

3. Scientific Approach and Results to Date

The main research thrust of this project was directed toward the development of theoretical techniques and tools needed for studying the dynamics and stability of proteins, the structure of nucleic acids, and the analysis of experimental data. With these tools in hand, theoretical calculations have been applied to experimental data from several Los Alamos research groups.

3.1. Development of Theoretical Tools for Biomolecular Structure and Dynamics Simulations

We have developed an efficient and robust code for performing simulations on large biomolecular systems in water-electrolyte solutions. In this code, we sought to implement the Ewald and the O(N) Adaptive Multipole algorithms for calculating electrostatic interactions in large systems. This code has been implemented in molecular dynamics or Monte Carlo simulations.

Collective variables have also been studied. Simulations of biomolecular dynamics are commonly interpreted in terms of harmonic or quasi-harmonic models for the dynamics of the system. These models assume that biomolecules exhibit oscillations around a single energy minimum. However, spectroscopic data on myoglobin suggest a broad distribution of energy barriers. This behavior has also been observed in other biomolecular systems. To elucidate the nature of protein dynamics, we studied a 1.2 nanosecond molecular dynamics trajectory of crambin in aqueous solution. This trajectory samples multiple local energy minima. Transitions between minima involve collective motions of amino acids over long distances. We have shown that nonlinear motions are responsible for most of the atomic fluctuations of the protein. These atomic fluctuations are not well described by large motions of individual atoms or a small group of atoms, but rather by concerted motions of many atoms. These nonlinear motions describe transitions between different basins of attraction. The signature of these motions manifests itself in local and global structural variables.

A method for extracting Molecule-Optimal Dynamic Coordinates (MODC) has been developed. A generalization of this method is used to identify small (1-3) dimensional subspaces of the configuration space describe the dynamics of the protein within the context of a nonlinear, multi-basin system [**1, 2**]. We have described the dynamics of biomolecules in

terms of an open Newtonian system (the protein) coupled to a stochastic system (the solvent). Auto correlation functions of the displacements along relevant MODC show that the protein loses memory of its configuration within a few picoseconds. The diffusion of the protein in configuration space is anomalous, namely, the time dependence of the mean square displacement is not proportional to time (t), but to t^{2HD} where 2HD is a nontrivial fractional exponent. Therefore, transitions among energy minima far apart in configuration space exhibit a stretched-exponential time dependence, which is found to scale as $t^{-2HD} e^{-t-2HD}$, with HD less than 0.5 [3]. This picture is consistent with a model suggested by Frauenfelder, *et al.* [4] and to explain multiple time scale relaxation processes observed in myoglobin.

A major development in simulation codes has been the implementation of the Fast Multipole Algorithm (FMA) on the CM-5 computer. This work was done in collaboration with LANL personnel and Norman Wagner (University of Delaware), who were developing a version based on a Cartesian tensor formulation, as compared to our version based on spherical harmonics. Both methods were successfully implemented, and it was demonstrated that our spherical harmonic version was more accurate and compact, as was expected. However, the Cartesian version was found to be faster, though the origin of this remains obscure. Collaborative work is continuing on optimizing this code, merging it with a molecular dynamics code, and adding biochemical features.

Another major area of work has been merging the Cray-YMP version of the code with a molecular dynamics code. This has involved adding coding to define the dimensionality of the problem at run time (i.e., number of atoms, number of boxes, and order of the expansion tensors). This also required the addition of an Ewald sum section so that infinite lattices can be treated properly. This was done, basically following the approach of Schmidt and Lee [5], though their method was limited to a cubic unit cell. This approach was generalized to the case for a triclinic cell, and implemented so that an arbitrary accuracy limit (up to machine precision) would be met. The timing for this modification was found to be incidental compared to the remainder of the code. After becoming aware of a Fourier-transform-based algorithm developed for accelerating the most time-consuming portion of the FMA (i.e., the transformation of the electrostatic multipole expansions into potential moment expansions) from order L^4 to order $L^2 \log(L)$, we have begun implementation of this into the Cray-YMP version of a molecular dynamics code, which will then serve as the basis for a PVM message-passing version of the code.

We have analyzed and implemented a continuum dielectric model for solvation of electrostatic interactions in aqueous solutions. The method, which was developed to solve the requisite macroscopic Poisson equation, is essentially a boundary element method. However, it differs importantly from previous numerical algorithms for this problem in sampling the

molecular surface utilizing quasi-random number series. This approach permits explicit demonstration of numerical convergence and systematic utilization of coarse-grained results. The method was tested by application to a range of solvation problems in aqueous solutions, including the potentials of mean force for (a) NaCl; (b) the S_N2-reactive system chloride-methyl chloride; and the S_N2-reactive system hydroxide-formaldehyde. The results of these tests were analyzed theoretically. A large part of the success of the dielectric model is likely associated with physically reasonable solvation stabilization of ionic fragments. The encouraging success of the dielectric model motivated us to consider further the molecular theory corresponding to the model. We obtained the first identification of that molecular theory. Analysis of that theory and testing with molecular simulation data provided a basic understanding of the conventional parametertizations of the dielectric model [**6, 7**].

The free energies of hydration of ions exhibit an approximately quadratic dependence on the ionic charge, as predicted by the Born model. We have analyzed this behavior using second-order perturbation theory. This analysis provided an effective method for calculating free energies from equilibrium computer simulations. The average and the fluctuations of the electrostatic potential at charge sites appeared as the first coefficients in a Taylor expansion of the free energy of charging. Combining data from different charging states allowed the calculation of free energy profiles as a function of ionic charge. The first two coefficients of the Taylor expansion were accurately calculated from equilibrium simulations; but they were affected by a string-system-size dependence. We applied corrections for these finite-size effects by using Ewald lattice summation and adding self interactions consistently. Results have been presented for a model ion with methane-like Lennard-Jones parameters in single-point-charge water. We found two closely quadratic regimes with different parameters for positive and negative ions. Negative ions are found to be more strongly solvated when compared to the positive ions of equal size, as measured by the solvation free energies. We ascribed this preference of negative ions to their strong interaction with water hydrogen atoms. We consistently found a positive electrostatic potential at the center of uncharged Lennard-Jones particles, which also favorably effects the free energy of solvation of negative ions.

The free energy of hydration of water has also been investigated. Using perturbation theory, we studied the chemical potential of water as a function of charge. By calculating the electrostatic energy fluctuations of two states (i.e., fully charged and uncharged), we were able to determine accurate values for the dependence of the chemical potential on charge. We found identical results for the chemical potential difference of fully charged and uncharged water from overlapping-histogram and acceptance-ratio methods and by smoothly connecting the curves of direct exponential averages. Our results agreed well with those published by Rich and Berne [8] with respect to both the chemical potential difference and its dependence on the charge

coupling parameter. We observed significant deviations from simple Gaussian fluctuation statistics. The dependence on the coupling parameter is not quadratic, as would be inferred from linear continuum methods of electrostatics. Two articles have been accepted for publication [**9, 10**].

3.2. Analysis of 2D-NMR Data, Analysis of X-Ray and Neutron Diffraction Data, Studies of Protein Structure and Dynamics, and Studies of Nucleic Acid Structure

We have developed a set of tools that integrate optimization algorithms, and molecular modeling techniques to analyze multi-dimensional NMR data. Molecular dynamics simulations are performed using our best simulation code to include the effect of local motions in the interpretation of NMR data in terms of structural constraints. This software is integrated with available graphics display software. The effect of base substitutions on the unusual DNA structures such as the stem-loop structures of the human centromeric DNA has been studied. Molecular modeling of the human centromere DNA repeat, $(AATGG)_4$, indicated that some local structural components are absolutely necessary for the stability of the structure. A combination of NMR measurements and calorimetric studies of various mutants has been done to test the theoretical predictions. Excellent qualitative agreement has been obtained [**11, 12**].

One of our goals in the protein dynamics experiments are to study the CO-Mb by Laue Diffraction at various temperatures and use these data to determine the temperature dependence of the Debye-Waller factors; and to obtain Fourier difference maps of the rebinding of CO to Mb after photolysis at low temperatures and on a time scale of 100 seconds. Another goal is the investigation of the motions that occur in Mb at short times after photo dissociation. Such motions have already been studied, but the existence at LANL of fast systems with detection in the infrared promises to yield new insight. Theoretical studies and neutron diffraction experiments are being conducted with the goal of determining the hydration structure of biomolecules.

Significant progress has been made in the analysis of the hydration of Mb. The hydrogen bond energy of water molecules, which are localized in the neutron map, were calculated and compared with water molecules from the dynamics simulation of Mb in solution. The dynamic simulation showed that few water molecules are permanently bound to the protein and most of the water molecules observed in the neutron diffraction analysis depict the average structure. A detailed analysis of the neutron maps shows that all of the well-bound water molecules form multiple hydrogen bonds. This dynamic analysis gives some insight into water hydration and explains the differences observed in protein hydration studies by

diffraction and NMR spectroscopy techniques. A paper describing this work has been published [13].

This study has been followed by a similar approach, but using a repetitive crystal lattice to study the effect of crystal packing constraints on the dynamics of the protein and the solvent. On completion of the x-ray and neutron diffraction investigation of perdeuterated myoglobin, the deuterated version will be studied in the same way to assess differences in structure and dynamics between these isotopically different forms. The use of deuterated proteins is of great advantage in neutron studies since it eliminates the large incoherent background produced by hydrogen atoms.

RNA molecules are involved directly in many different biological functions. How RNA molecules fold into their functional three-dimensional structures is an important question to address. Compared to protein and DNA structure, little is known about the atomic structure of RNA. Only twenty RNA structures exist in the Protein Data Bank. By comparison, approximately two hundred DNA structures and approximately three thousand protein structures are found in the in Protein Data Bank. The relatively large sizes of the molecules make it particularly difficult in solving the tertiary structures of RNA molecules, either experimentally or computationally. The modeling of the tertiary structure of RNA is a very complex and difficult task. Recent developments in the synthesis and sorting of RNA molecules, and searches of large sets of sequences can lead to molecules with extremely high specificity to almost any ligand. By exploiting these techniques, RNA molecules can be designed such that specific ligands can be targeted. This has open a new field in the development of drugs.

Experimental evidence indicating contacts between residues, which are not related by their secondary structure, exist for 5S and 16S RNA molecules; and in general, can be easily obtained for any RNA molecule. Based on a set of complete reduced coordinates and a general knowledge of the nucleic acid's secondary structure and low resolution tertiary contacts, an approach to model RNA tertiary structures at atomic resolution has been developed. Information about tertiary contacts between the bulge and loop regions [14] allow us to model this section of the molecule by three helices, three loops (two loops three bases long and one loop five bases long) and three phosphate linkages. The resultant structure is closely packed. This two-step approach has been tested successfully in folding of a pseudo knot motif (bases 500-545) in E. *coli* 16 S RNA [15]. The atomic structure of the binding domain of a RNA molecule that binds the bronchodilator theophylline with high affinity and specificity was also predicted using this two-step approach [16].

References

[1] Garcia, A.E.; Soumpasis, D.M.; and Jovin, T.M., "Dynamics and Relative Stability of Parallel and Anti-parallel Stranded DNA Duplexes," *Biophys. J.*, **66**, 1742-1755 (1994).

[2] Garcia, A.E., "Multi-basin Dynamics of Protein in Aqueous Solution". in Nonlinear Excitation in Biomolecules, M. Peyrard, Editor. Pringer Verlag, New York. pp. 191-206.

[3] Garcia, A.E.; Blumemfeld, R.; Hummer, G.; and Soberhart, J. , "Diffusion of a Protein in Configurational Space", in Proceedings of the Ninth Conversation, R.H. Sarma and M. H. Sarma, Editors. Adenine Press, NY. 1995.

[4] Fraunfelder, H. ; Sligar, S.G.; and Wolynes, P.G., *Science* , **254**, 1598-1603.

[5] Schmidt and Lee , *J. Stat. Phys.*, **63**, 1223, (1991).

[6] Pratt, L.R.; Hummer, G.; and Garcia, A.E., ``Ion Pair Potentials-of-Mean-Force in Water," *Biophys.* **51**, 147-165 (1994). LA-UR-93-4205. chem-ph@xxx.lanl.gov e-print server paper #9404001.

[7] Tawa , G.J. and Pratt, L.R. ``Tests of Dielectric Model Descriptions of Chemical Charge Displacements in Water," preprint 1994 for ACS symposium volume ``Structure, Energetics, and Reactivity in Aqueous Solution," edited by C. J. Cramer and D. G. Truhlar. LA-UR-94-431. chem-ph@xxx.lanl.gov e-print server paper #9404002.

[8] Rich and Berne *J. Am. Chem. Soc.*, **116**, 3949, (1994)

[9] Hummer, G.; Pratt, L.R.; and Garcia, A.E., "The Hydration Free Energy of Water," *J. Phys. Chem.*, **99**, 14188-14194. LAUR-95-1612

[10] Hummer, G; Pratt, L.R.; and Garcia, A.E., "On the Free Energy of Ionic Hydration," *J. Phys. Chem.* (in press). LAUR-95-1161.

[11.] Catasti, P. ; Gupta, G.; Garcia, A.E.; Ratliff, R.; Hong, L.; Yau, P.; Moyziz, R.; and Bradburry, E.M. *Biochemistry,* **33**, 3819-3830 (1994).

[12] Gupta, G.; Garcia, A.E.; Catasti, P.; Ratliff, R.; Bradburry, E.M.; and Moyzis, R.K., in Proceedings of the Eighth Conversation, R.H. Sarma and M.H. Sarma, Editors,. pp. 137-154 (1994).

[13] Gu, W. and Schoenborn, B.P., *Proteins Struc. Fun.& Gen.,***22,** 22-26, (1995).

[14] Guetell, *et al.*, *Microbiological Review*, **58**, 10-26, (1994)

[15] Tung, C.-S. "A Geometrical Approach in Folding a Pseudoknot Motif within the E *coli* 16S-RNA," LA-UR: 94-3673, *Biopolymer*, submitted, (1995).

[16] Tung, C.-S.; Oprea, T. I.; Hummer, G.; and Garcia, A. E. "Three-Dimensional Model of a Selective Theophylline-Binding RNA Molecule," LA-UR:95-1798, *J. Mol. Recog.*, submitted, (1995).