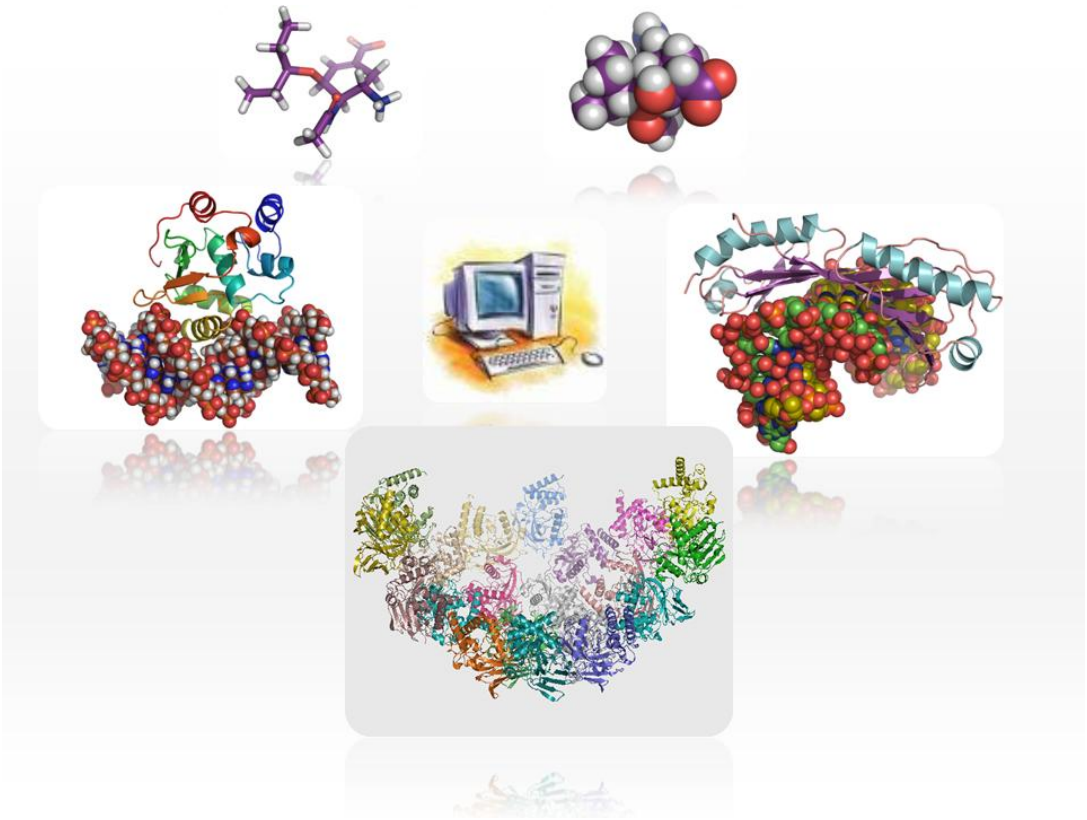

RESEARCH STATEMENT AND PLAN

(FROM BASIC BUILDING BLOCKS TO MACHINE LEARNING APPROACHES)

Prepared by: Hiqmet Kamberaj, Ph.D.

Date: August 2018

RESEARCH STATEMENT

Macromolecules (such as, proteins, DNA and their complexes) are characterized by complex topologies of the energy landscapes and a large number of energy minima. Moreover, their dynamics involves motions of different time and size scales. The biological function of proteins is generally determined by correlated fluctuations involving large portions of the structure. These dominant correlated motions in macromolecules will certainly influence their behaviors on micro and nano-scales, and must be understood and appropriately handled in order to enable rational protein-based technologies.

Therefore, developing advanced computational methods that are capable to determine efficiently the equilibrium conformations of macromolecules in aqueous solvent is a major goal of computational biophysics. Molecular dynamics and/or Monte Carlo methods are commonly used to drive such systems and have provided a significant contribution to understanding of protein structure and function.

To overcome the problems arising from the time and size scale limitations of simulations of such systems, the molecular simulations methods, which allow simulations of systems of biologically relevant size and time scale, have shown great interest. The main aim is to further develop these methods that would allow investigating and predicting the structure and function of biological systems, as well as design rules for related technologies.

SUMMARY OF RESEARCH PLAN

Objectives

The main objectives of my research activity include the understanding of structure and function of biological systems and designing rules for related technologies by overcoming problems arising from the time and size scale limitations of these systems.

Goals

The main goals of my research plan include development of novel methods for improving conformational sampling efficiency of molecular dynamics techniques, enhancing transition path sampling for rare event simulations, information flow in the biological systems using information theoretic measures, measurement of high-order correlated motions in complex

dynamical systems, prediction of free energies using Machine Learning approaches combined with physical models and numerical modeling techniques for use in molecular dynamics method.

Solution

Three strategies will be pursued in enhancing conformational sampling and transition path sampling of rare events in molecular dynamics simulations of complex molecular systems: **(1)** Development of replica exchange algorithms for use in molecular dynamics and/or Monte Carlo simulations and improving the stability of numerical methods for solving differential equations by use of symplectic numerical integrators; **(2)** Enhancing conformational sampling and transition path sampling of molecular dynamics simulations using biasing methods either on potential energy function or on the equations of motion in order to overcome the problems of time scale limitations; **(3)** Development and use of coarse-grained models for bimolecular systems by using renormalization group theory for overcoming the problems of time and size scale limitations.

Two measures of the information flow will be introduced and evaluated on real dynamical systems: **(1)** *Symbolic transfer entropy* and **(2)** *local symbolic transfer entropy* using the information theoretic measures. These two methods will be used to build up global and local network of interactions in complex (bio)molecular systems by using clustering techniques in order to understand the *communication signal pathways* in the systems involving interactions within and between proteins. Attempts to relate the local symbolic transfer entropy with thermodynamics measures (such as heat flow) will be made, in particular, for studying state transitions and biological system stability. Reduction of the dimensionality of the system features (either at amino acid level, such as when determining the communication signal pathways within proteins or at protein-protein and protein-DNA/RNA interfaces, or at protein level, such as when determining signal transduction pathway steps) will be investigated using Machine Learning approaches, such as Artificial Neural Networks combined with encoding-decoding algorithms.

Two methods will also be introduced and compared with each other for measurement of high-order correlated motions in complex dynamical systems: **(1)** Continuous mutual information and **(2)** symbolic mutual information. The development of these two methods will be based on the information theory.

Project Outline

To accomplish the above mentioned research activities it is thought to combine theory and computer simulations. Typically, molecular dynamics methods will be used along with

information theoretic measures and causalities. Based on the nature of each of the above mentioned problem, we can distinguish four different research projects:

1) *Enhancing conformational sampling of molecular dynamics simulations and transition path sampling of rare events in (bio)molecular systems.* In order to accomplish this project, method proposed in Ref.[1] will be implemented in CHARMM program.[2] This method combines standard molecular dynamics and swarm particle intelligence[3] methods to enhance conformational sampling. The new method will be applied to investigate large conformational changes in proteins and to study transition path sampling of rare events. First, the method will be applied to known test problems, then to more complex (bio)molecular systems.[4] Other methods of enhanced sampling techniques will be investigated as well.[5]

2) *Global and local symbolic transfer entropies as measures of the information flow in complex dynamical systems useful tools for constructing complex interaction networks and investigating heat flow of state transitions and their stability.* A new method will be introduced for calculation of the local symbolic transfer entropy, which is based on the symbolic transfer entropy method introduced elsewhere,[6] which is already implemented in CHARMM program and using a newly established software.[7] In this study, two problems will be the main goals: **(1)** Building global and local network of interactions using these two measures, and **(2)** relating local symbolic transfer entropy with

[1]H. Kamberaj, Conformational Sampling Enhancement of Replica Exchange Molecular Dynamics Simulations Using Swarm Particle Intelligence, *The Journal of Chemical Physics*, 143, 124105, 2015.

[2]Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry,* 30, 1545, 2009.

[3]H. Kamberaj, Q-Gaussian Swarm Quantum Particle Intelligence on Predicting Global Minimum of Potential Energy Function, *Applied Mathematics and Computation*, 229, 94, 2014.

[4] H. Kamberaj, (2018) Faster Protein Folding Using Enhanced Conformational Sampling of Molecular Dynamics Simulation, *Journal of Molecular Graphics and Modelling*, 81, 32-49.

[5]J. Spiriti, H. Kamberaj and A. van der Vaart, Development and application of enhanced sampling techniques to simulate the long-time scale dynamics of biomolecular systems, International Journal of Quantum Chemistry, 112, 33, 2012.

[6]H. Kamberaj and A. van der Vaart, Extracting the causality of correlated motions from molecular dynamics simulations, Biophysical Journal, 97, 1747, 2009.

[7] Dh. Nebiu and H. Kamberaj, Symbolic Information Flow Measurement (SIFM): A Software for Measurement of Information Flow Using Symbolic Analysis, to be submitted in 2018.

thermodynamics properties of the system.[8] In particular, the main aim will be to thermodynamically associate the increase in the local transfer entropy with the stability of a system (such as biological system, for instance, the protein stability). Other goal of this project will include prediction of the state transitions of a system based on the local transfer entropy measure, by relating transfer entropy with heat flow to the system (characterized by the so-called *collective coordinates*) from the *environment* defined in the context of the other faster degrees of freedom of the system.[9] Possibilities of the use of these two methods to other problems of interests involving dynamical systems characterized by differential equations will also be considered. Reduction of the dimensionality of the system features, in order to determine the collective coordinates, will be investigated using Machine Learning approaches, such as Artificial Neural Networks combined with encoding-decoding algorithms.[10]

3) *Development and use of symplectic numerical integrators for second order differential equations characterizing dynamics of (bio)molecular systems based on Hamiltonian formalism.*[11] Long MD simulations (from hundreds of microseconds to milliseconds time scale) have been achieved by computer engineering using standard MD method or by enhancing the rate of barrier crossing events either without introducing a bias, or introducing bias that can be rigorously removed a posteriori.[12] [13]Other methods have also been developed based on the multiple time step integrators, such as reference system propagator algorithm.[14] In this research project, Hamiltonian formalism will be used to generate Hamiltonian dynamical systems (preserving symplectic structure), which can then be combined with symplectic numerical integrators to generate stable numerical integrators with larger time steps.[15]

---

[8]Mikhail Prokopenko and Joseph T. Lizier, Transfer Entropy and Transient Limits of Computation, Scientific Reports, 4, 5394, 2014.

[9]H. Kamberaj, A Theoretical Model for the Collective Motion of Proteins by Means of Principal Component Analysis, Central European Journal of Physics, 9, 96, 2011.

[10] Dh. Nebiu and H. Kamberaj, Symbolic Information Flow Measurement (SIFM): A Software for Measurement of Information Flow Using Symbolic Analysis, to be submitted in 2018.

[11] H. Kamberaj, Advanced Methods used in molecular dynamics simulation of macromolecules, in Advanced Computational and Applied Engineering Research, Nova Science Publishers, Inc., Submitted 2018.

[12] H. Kamberaj, (2018) Faster Protein Folding Using Enhanced Conformational Sampling of Molecular Dynamics Simulation, *Journal of Molecular Graphics and Modelling*, 81, 32-49.

[13] H. Kamberaj, (2015) Conformational Sampling Enhancement of Replica Exchange Molecular Dynamics Simulations Using Swarm Particle Intelligence, *The Journal of Chemical Physics*, 143, 124105.

[14]M.E.Tuckerman, B.J.Berne, and G.J.Martyna. The *Journal of Chemical Physics,* 97,1990, 1992.

[15]H. Kamberaj, R.J. Low and M.P. Neal, Symplectic and time reversible integrators for molecular dynamics simulations of rigid molecules, *The Journal of Chemical Physics*, 122, 224114, 2005.

4) *Development of coarse-grained models for bimolecular systems based on the renormalization group theory*. To overcome problems arising from the time and size scale limitations of complex molecular systems the coarse-grained models have also shown a great interest. Here, we will try to use the renormalization group theory to develop new coarse-grained models for proteins, which can then be used in molecular dynamics and/or Monte Carlo simulations.

5) *Development solvation models for prediction of absolute and relative binding free energies of protein complexes in solution*. To overcome problems of computation efforts, solvation continuum models[16] will be combined with appropriate Machine Learning approaches to predict solvation free energies using as benchmarks the available datasets and/or molecular dynamics simulations combined with experimental results of mutations.[17]

Project time line

For each project, Table 1 shows the time line for a typically 36 months (3 years) research plan, which is composed into three time periods:

(1) First time period includes writing of the research proposals for a particular project, which involves: project study, literature review, database search and software proposals. This part will ideally take up to 12 months.

(1) During this period several undergraduate and/or graduate research projects can be offered for testing the software and/or models to simple physical systems.

(2) In the next time period, the project team will be formed by offering research project proposals to PhD students. This step will also include extending collaboration team with (inter)national partners (new and existing one.) This part will ideally take up to 6 months.

(1) During this period final year undergraduate and graduate thesis projects can be offered to more advanced students to verify computational models to more complex physical systems.

[16] R. Izairi and H. Kamberaj, (2017) Comparison Study of Polar and Non-polar Contributions to Solvation Free Energy, *Journal of Chemical Information and Modeling*, 57(10), 2539-2553.

[17] H. Kamberaj, (to be submitted in October 2018 as Special Issue), Prediction of Solvation Free Energy Using a Bootstrapping Swarm Artificial Neural Network Method: A Machine Learning Approach, *Journal of Chemical Information and Modelling*.

(3) The third time period will include research project accomplishment: data acquisition, data analysis, report writing and publications in peer-reviewed journals. This part will ideally take up to 18 months.

(1) Also, during this period several undergraduate and graduate student projects can be offered writing and testing models for statistical data analysis.

	Time Line (Months)		
	12	6	18
Create a proposal			
Form the project team (Add (inter) national partners)			
Project accomplishment			

Table 1: The project time line.