論文内容の要旨
論文題目：
生体分子運動の縮約表現
(Reduced Ensemble Representation of Biomolecular System)

氏名　：　櫻庭 俊

Dynamics of the biomolecules are complex, which enables the complex functionality of the biomolecular systems. One of the best ways to understand such complex systems is, simply, seeing how it works. Molecular simulation enables us to make a "sandbox" of the proteins inside computers, with which we researchers can observe its intrinsic motions or its response to changes in the system; we can see how it works. Combined with other experimental techniques, it nowadays became a mature and an essential tool for protein science.

One of the problems with the molecular simulation is how we understand the results. Proteins are high dimensional systems, and the simulations of them generate even tera-bytes sized data within a week. The ways to extract data from the enormous data space is necessary, and the result of extraction has to be human-understandable. Herein, I propose two approaches to cope with the problem. One is a method to estimate the *stability* of the protein's substates, without requiring long sequential run. The other is a method to determine the best axes to represent protein's dynamics, or *reaction coordinates* of the proteins, automatically from the simulation.

1 Combining Multiple Non-equilibrium Dynamics Simulation Based on Markov Model

In the molecular simulation, the snapshots of the system are taken with interval, which consists of trajectory. Trajectory is considered as the representative set for the molecule. However, even with today's hardware and simulation algorithm, there are still a large gap between the time-frame of simulation and that

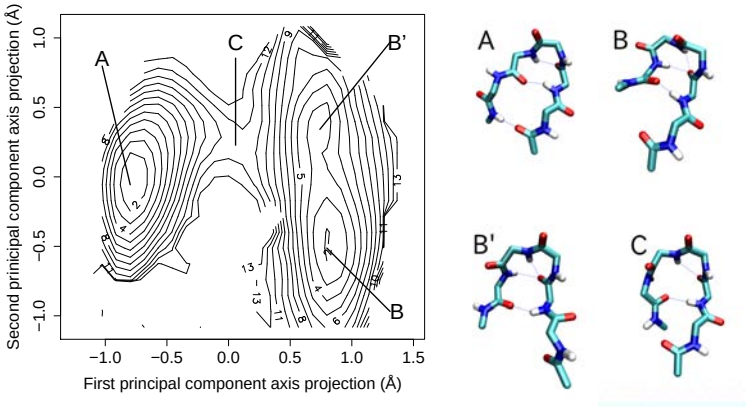

Figure 1: Energy landscape of 5-residue peptide Met-enkephalin, obtained with MMMM. Representative structures for each characteristic point is also shown on the right side.

of protein's functional motion, such as enzyme activity or receptor binding. Possible workaround is to improve sampling with modified dynamics. These methods, usually called generalized ensemble, sample snapshots from simulations with modified dynamics; after obtaining data, each snapshot of the trajectory is weighed according to the magnitude of modification. The pair of trajectory and weigh represents the protein's character. One of the drawbacks of these methods is that methods rely on the equivalence of statistical ensemble average and time average. Because of this, for the large systems these methods require long equilibration time before researchers can start sampling snapshots. Thus generalized ensemble is at this moment impractical for the biomolecules.

In this research I pursued the new analysis method which calculates the statistical ensemble of the system, but which does not require total equilibration of the system. Starting from the assumption of Markovity, the dynamics of the system can be represented by a transition matrix M. The equilibrium probability density $\boldsymbol{\pi}$ can be obtaind by the left diagonalization of the M, since $\boldsymbol{\pi}$ is the left eigenvector of M corresponding to eigenvalue 1.

$$\boldsymbol{\pi} = M\boldsymbol{\pi}. \tag{1}$$

Simple form to compute equilibrium probability distribution was therefore obtained. Extending this form to combine multiple simulations, I developed a method called Multiple Markovian transition Matrix Method (MMMM), based on the error analysis of the eigenproblems. MMMM was compared with other state-of-art ensemble value determination methods, and it showed better result with the case that the equilibration cannot be expected. Also, the method was tested with peptide energy landscape as an application, giving proper transition state structure. With these results, I showed the method is applicable to even a practical case.

2 Decomposing Protein into Components with Correlations

In section 1, reaction coordinates were considered to be given, but finding proper reaction coordinates is a non-trivial task. Also, because of the curse of dimensionality, above problem does not scale well. The number of states blows-up with the increase of dimension. One of the reasons is that proteins consist of strongly coupled components. If there are many uncoupled components in a protein, each component can be analyzed in divide-and-conquer fashion. In reality, it is non-trivial task to find uncoupled components.

In order to identify the correlated modes and to decompose proteins into uncoupled components, I borrowed an idea from the field of the signal processing. In this thesis Independent subspace analysis (ISA) method is introduced to the concept of the biomolecular collective motion. A linear projection of the coordinates $\boldsymbol{x}$ is considered:

$$\boldsymbol{s}(t) = A\boldsymbol{x}(t). \tag{2}$$

With ISA, projected vector $\boldsymbol{s}$ and its rotation matrix A can be determined so as to minimize the number of strongly correlated modes. Procedure based on subspace joint approximate diagonalization of eigenmatrices (SJADE) algorithm is employed to perform ISA. ISA/SJADE dissects multiple dimensions into irreducible "blocks", in which projection to each dimension have strong correlation within the same block, while there are no or very small correlation over different blocks. With ISA/SJADE, the result of 100-ns MD simulation of T4 lysozyme was analyzed for testing purpose. Result showed the modes determined from ISA explain long-ranged correlation of modes, and it successfully found the modes which are strongly correlated to functionally important residues reported from mutation experiments. From these results, ISA is shown to be powerful technique for analyzing protein dynamics.

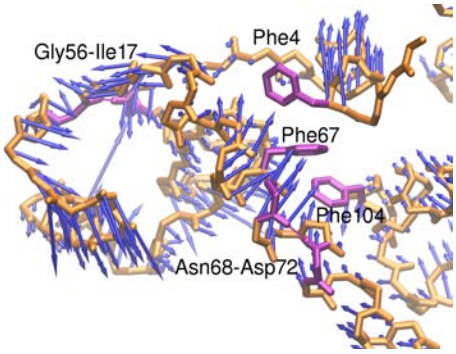

Figure 2: Arrow representation of SJADE mode 5, and the residues which strongly correlated to mode 5. Presented residues are confirmed to be strongly correlated to the motion of the mode 5.