Copyright

Tianwu Zang

2014
RICE UNIVERSITY

Enhanced Sampling Method in Statistical Physics and
Large-Scale Molecular Simulation of Complex Systems

by

Tianwu Zang

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE

Jianpeng Ma, Chair
Professor of Bioengineering, Rice University
and Lodwick T. Bolin Professor of
Biochemistry, Baylor College of Medicine

Ching-Hwa Kiang,
Associate Professor of Physics & Astronomy
and Bioengineering

Robert M. Raphael
Associate Professor of Bioengineering

HOUSTON, TEXAS
January 2014
ABSTRACT

Enhanced Sampling Method in Statistical Physics and Large-Scale Molecular Simulation of Complex Systems

by

Tianwu Zang

In large-scale complex systems, traditional computational methods in equilibrium statistical mechanics such as Monte Carlo simulation and molecular dynamics in canonical ensemble often face the broken ergodicity issue, which highly affects the performance and accuracy of simulation. The past decades have witnessed the development of generalized ensemble, which has significantly enhanced the efficiency of molecular simulation. In this thesis, we review typical generalized ensembles, such as multi-canonical ensemble, parallel tempering, simulating tempering and continuous simulated tempering (CST).

We also present a method called parallel continuous simulated tempering (PCST) for enhanced sampling in studying large complex systems. It mainly inherits and CST method in previous studies, while adopts the spirit of parallel tempering, by employing multiple copies with different temperature distributions. The sampling efficiency of PCST is tested in two-dimensional Ising model, Lennard-Jones liquid and all-atom folding simulation of a small globular protein trp-cage in explicit solvent. The results demonstrate that the PCST method has significantly improved
sampling efficiency compared with other methods and it is particularly effective in simulating systems with long relaxation time or correlation time.
Acknowledgments

I would like to thank my advisor, Dr. Jianpeng Ma, for his patience, enthusiasm and continuous support of my study and research. As one of the best and smartest scientists, Dr. Ma always points out the keys to the problem which guide the research direction to a correct way. Without his enlightenment, the new ideas in this thesis would never come into reality.

My sincere thanks also go to the rest of my thesis committee: Dr. Kiang and Dr. Raphael for their encouragement and insightful advices.

I would also like to thank all of my current and former colleagues in Dr.Ma’s lab, Cheng Zhang, Chong Zhang, Dong Chen, Hailiang Zhang, Jia Zeng, Kit Menlove, Linglin Yu, Mingyang Lu, Qian Mei, Yunxiang Liao, Yushao Cheng and Zhenwei Luo, for all the stimulating discussions and enjoyable time together.

Part of the work in my thesis was supported by Blue BioU at Rice University. I would like to thank Dr. Qiyou Jiang, who is a HPC expert at Rice, for his helpfulness and kindness in answering my every question related to research computing. I also thank Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC and grid resources.

Last but most important, I would like to thank my parents. Without their love and support, this thesis would have never been possible.
Contents

Acknowledgments .................................................................................................................. iv

Contents ................................................................................................................................... v

List of Figures ............................................................................................................................ vii

List of Equations ......................................................................................................................... viii

Basic Principles in Statistical Mechanics ............................................................................... 1
  1.1. Introduction ....................................................................................................................... 1
  1.2. Ensembles ........................................................................................................................ 2
    1.2.1. Microcanonical Ensemble ......................................................................................... 3
    1.2.2. Canonical Ensemble .................................................................................................. 4
  1.3. Relationships with Classical Thermodynamics ............................................................. 5

Computational Methods in Statistical Mechanics ................................................................. 8
  2.1. Constant Temperature (N, V, T) Simulation ................................................................. 8
    2.1.1. Monte Carlo Method .............................................................................................. 8
    2.1.2. Molecular Dynamics .............................................................................................. 9
    2.1.3. Limitations of NVT Simulation ............................................................................... 12
  2.2. Generalized Ensemble .................................................................................................... 13
    2.2.1. Introduction ............................................................................................................ 13
    2.2.2. Multi-canonical Ensemble ..................................................................................... 14
    2.2.3. Simulated Tempering ............................................................................................. 14
    2.2.4. Parallel Tempering .................................................................................................. 17

Advanced Sampling Techniques for Complex Systems ....................................................... 19
  3.1. Continuous Simulated Tempering (CST) ...................................................................... 19
  3.2. Parallel Continuous Simulated Tempering (PCST) ....................................................... 22
  3.3. Estimating Physical Properties in Simulated Tempering ............................................. 25
    3.3.1. Introduction .......................................................................................................... 25
    3.3.2. Adaptive Averaging Techniques ............................................................................. 26
    3.3.3. Multi-bin Estimator based on Integral Identities (MEII) ...................................... 27
  3.4. Simulation Protocols of PCST ......................................................................................... 31
3.5. Applications ........................................................................................................................................ 32
  3.5.1. Phase Transition Simulation in 2-dimensional Ising Model ....................................................... 32
  3.5.2. Lennard-Jones Fluid ..................................................................................................................... 34
  3.5.3. Folding Small Proteins in All-atom Model .................................................................................. 36

3.6. Discussions ......................................................................................................................................... 49

Conclusion .................................................................................................................................................. 51

References .................................................................................................................................................. 53
List of Figures

Figure 1 Calculated heat capacity using PCST and its absolute error. ..............33
Figure 2 Generated temperature distribution ....................................................34
Figure 3 The convergence of energy per atom in different methods .................36
Figure 4 $C_\alpha$ - RMSD from the native state in five trajectories using CST .........39
Figure 5 $C_\alpha$ - RMSD from the native state in five trajectories using PCST ..........40
Figure 6 RMSD-Rg relationship of $C_\alpha$ atoms in PCST simulation ..................42
Figure 7 The $C_\alpha$ - RMSD distribution in PCST method and CST method ............44
Figure 8 Enlargement of parts of traj 3 and traj 5 in Figure 5 ............................45
Figure 9 Molecular structure of two substates ..............................................46
Figure 10 Normalized joint distribution of the $C_\alpha$ (helix) - RMSD and $C_\alpha$ - RMSD of trp-cage in PCST simulations .........................................................48
List of Equations

Equation 1-1 Time average of measure physical quantities ........................................ 2
Equation 1-2 Hamilton’s equations in classical mechanics ........................................ 4
Equation 1-3 Boltzmann distribution ........................................................................ 4
Equation 1-4 Ensemble average of physical property A ............................................. 5
Equation 1-5 Microcanonical entropy ....................................................................... 5
Equation 1-6 Canonical average of physical property A .......................................... 6
Equation 1-7 Energy fluctuation in canonical ensemble ............................................. 6
Equation 1-8 Canonical entropy .............................................................................. 6
Equation 2-1 Acceptance ratio in Metropolis algorithm ............................................ 9
Equation 2-2 Newton’s equation guiding molecular dynamics .................................. 10
Equation 2-3 Leap-frog integrator in molecular dynamics ........................................ 10
Equation 2-4 Kinetic energy random walk in velocity-rescaling thermostat .......... 11
Equation 2-5 Configuraional probability distribution in simulated tempering .... 15
Equation 2-6 Acceptance ratio for temperature move in simulated tempering .... 15
Equation 2-7 Configuraional probability distribution in parallel tempering .......... 17
Equation 2-8 Acceptance ratio for replica exchange in parallel tempering .......... 17
Equation 3-1 Probability distribution of configuration in CST .................................. 19
Equation 3-2 General form of one-dimentional Fokker-Planck equation .............. 20
Equation 3-3 Random walk in temperature space ..................................................... 21
Equation 3-4 Random walk in temperature space ..................................................... 21
Equation 3-5 Probability distribution of configuration in PCST .................................. 23
Equation 3-6 Random walk in temperature space in PCST ........................................... 23
Equation 3-7 Corresponding Langevin equation of Equation 3-6 ................................. 23
Equation 3-8 Acceptance ratio for parameter exchange in PCST ................................. 24
Equation 3-9 Acceptance ratio in PCST for same variances ........................................ 24
Equation 3-10 Langevin equation after introducing multiple bins ............................... 26
Equation 3-11 Weighted mean and variance in each small bin .................................. 27
Equation 3-12 Incremental calculation of weighted mean and variance ..................... 27
Equation 3-13 Relationship between heat capacity and energy fluctuation .................. 28
Equation 3-14 Integral identity for any continuous function ........................................ 28
Equation 3-15 Integral identity in Equation 3-14 in vector form .................................. 28
Equation 3-16 Unbiased estimation of average energy in each small bin ...................... 29
Equation 3-17 Mathematical form of $\phi(\beta)$ .......................................................... 29
Equation 3-18 Specific value of $\phi$ and $\psi$ in each small bin ...................................... 30
1.1. Introduction

Statistical mechanics serves as a bridge connecting the macroscopic and microscopic world. It tells that the macroscopic properties in classical thermodynamics, such as pressure and internal energy, can be obtained by observing and doing statistics on microscopic states.

Although it seems to be very quick to measure a physical quantity in our macroscopic world, the system has already experienced countless number of times jumping from one microscopic state to another during the experiment. Therefore, the measured quantities are actually the average of all the quantities corresponding to the experienced microscopic states over time, i.e.
\[ \langle A \rangle_t = \frac{1}{T} \int_0^T A(t) \, dt \]

**Equation 1-1 Time average of measure physical quantities**

where \( T \) represents the time of the measurement. If \( T \) is sufficiently long, \( \langle A \rangle_t \) no longer changes with time and converges to a fixed value. This value is the macroscopic value of \( A \).

In the case above, if we focus on the frequency \( w_i \) of any microscopic state \( i \), we can find that \( w_i \) does not change with time as well. Therefore, we can imagine a type of ideal experiment under the following conditions: instant measurements are taken on independent systems with the same macroscopic conditions. The observations corresponding to different microstates can be regarded as “snapshots” from the trajectory of the system in the phase space. We call the collection of these “snapshots” the statistical ensemble of the system. In equilibrium, the probability for a microstate \( i \) in this ensemble should be equal to the frequency \( w_i \) mentioned above. The equivalence of ensemble average and time average of physical properties in equilibrium systems is one of the major principles in statistical mechanics.

**1.2. Ensembles**

An ensemble is a collection of microscopic states under certain weights. There are different kinds of ensembles in statistical mechanics, e.g. microcanonical ensemble, canonical ensemble, grand canonical ensemble and isothermal–isobaric
ensemble, etc. We will briefly summarize the micocanonical ensemble and canonical ensemble, which are the two major ensembles related to this thesis, in the following sections.

1.2.1. **Microcanonical Ensemble**

Microcanonical ensemble, or \((N, V, E)\) ensemble, depicts the evolution of an isolated system. Completely enclosed by walls, the isolated system does not exchange any matter or heat with the outside world, and the Hamiltonian of the system is conserved over time.

A fundamental postulate, which is often called *equal a priori probability postulate*\(^1\), states that

> *For an isolated system with an exactly known energy and exactly known composition, the system can be found with equal probability in any microstate consistent with that knowledge.*

This postulate brings the equal weight of any microstates in microcanonical ensemble, i.e. \(w_i = 1\).

In classical mechanics, the position and momentum \(\{\bar{p}, \bar{q}\}\) of the system evolve under the Hamilton’s equations

\[
\begin{aligned}
\dot{\bar{p}} &= -\frac{\partial H}{\partial \bar{q}} \\
\dot{\bar{q}} &= \frac{\partial H}{\partial \bar{p}}
\end{aligned}
\]
Equation 1-2 Hamilton’s equations in classical mechanics

The Hamilton’s equations are deterministic, and it describes the dynamical behaviors of the system on the \((N - 1)\)-dimensional energy surface in the phase space. Given a fixed Hamiltonian, we can conclude that there exists only one trajectory passing any point on this surface. This trajectory goes through every point on the energy surface, and the average time for the system to stay in any microscopic state is equal to any other microscopic states on this surface.

1.2.2. Canonical Ensemble

Different from microcanonical ensemble, canonical ensemble describes the system that is imagined to be in thermal equilibrium with an infinite-large heat bath. Under the second law of thermodynamics, the probability for the system to be in the microscopic \(i\) is only related to the energy of this state \(E_i\). The weight of this state \(w_i\) follows the Boltzmann distribution with the mathematical form

\[
w_i \propto e^{-\beta E_i}.
\]

Equation 1-3 Boltzmann distribution

where \(\beta \equiv 1/k_B T\) is the reciprocal temperature.
1.3. Relationships with Classical Thermodynamics

As is discussed before that different ensembles use different weights of microscopic states, in the phase space. Therefore, it is useful to define the partition function \( Z = \sum_i w_i \) which represents the sum of the weights of all the microstates in this ensemble. Therefore, the ensemble average of a physical property \( A \) can always be written as

\[
\langle A \rangle = \frac{\sum_i w_i A_i}{Z} = \sum_i p_i A_i,
\]

**Equation 1-4 Ensemble average of physical property \( A \)**

where \( p_i \) denotes the probability for the system to stay in the microscopic state \( i \).

In micocanonical ensemble, \( w_i = 1 \), which means that the partition function can be defined as the number of states in the \( (E, E + \Delta E) \) shell in the phase space. This function is often called the density of states \( \Omega(E) \), which is only related to the energy \( E \). Based on the definition of Gibbs entropy, the microcanonical entropy can be obtained by

\[
S = -k_B \sum_i p_i \ln p_i = -k_B \sum_i \frac{1}{\Omega(E)} \ln \frac{1}{\Omega(E)} = k_B \ln \Omega(E),
\]

**Equation 1-5 Microcanonical entropy**

which is also called the Boltzmann equation.
In canonical ensemble, we can write the ensemble average of physical property $A$ under Boltzmann weights as

$$
\langle A \rangle = \frac{\sum_i e^{-\beta E_i} A_i}{Z}.
$$

Equation 1-6 Canonical average of physical property $A$

Particularly, if $A$ represents the total energy, then the equation above can be simplified as $\langle E \rangle = -\partial_\beta \ln Z$.

We can also calculate many other physical properties by simply using the partition function in canonical ensemble. For example, the fluctuation of energy in canonical ensemble is

$$
\langle (\Delta E)^2 \rangle \equiv \langle E^2 \rangle - \langle E \rangle^2 = \frac{\partial^2 \ln Z}{\partial \beta^2} = C_v k_B T^2,
$$

Equation 1-7 Energy fluctuation in canonical ensemble

where $C_v$ represents the heat capacity and $T = 1/k_B \beta$ is the temperature. The canonical entropy is

$$
S = -k_B \sum_i e^{-\beta E_i} \ln(e^{-\beta E_i}) = \frac{\langle E \rangle}{T} + k_B \ln Z = \frac{\partial (k_B T \ln Z)}{\partial T} = -\frac{\partial F}{\partial T},
$$

Equation 1-8 Canonical entropy
where $F \equiv -k_B T \ln Z$ is defined as the Helmholtz free energy.

As many physical properties can be calculated via partition functions, estimating partition functions is one of the major tasks in equilibrium statistical physics. On the other hand, there are only very few systems in which partition function can be solved analytically, such as 1-dimensional and 2-dimensional Ising models, ideal gas, etc. For the rest of the systems, the ensemble averages of physical properties as well as the partition function have to be determined by either experimental or computational approaches.

Compared with the experimental methods, computational methods are more economic and less time-demanding. However, only the dynamics of systems with small size in a short timescale are allowed for simulation on current computers, which puts demands on developing advanced computational methods for simulating events with longer timescale and estimating physical properties accurately for larger systems.
2.1. Constant Temperature \((N, V, T)\) Simulation

2.1.1. Monte Carlo Method

In canonical ensemble, the ensemble average of any physical property \((A)\) is \(\int A(X) P_{\text{Boltz}}(X; \beta) dX\), where \(P_{\text{Boltz}}(X; \beta)\) represents the Boltzmann distribution at temperature \(\beta\) with the mathematical form of \(e^{-\beta E(X)} / \int e^{-\beta E(X)} dX\). Therefore, to compute the ensemble average in canonical ensemble is essentially a high-dimensional numerical integration problem.

Proposed to solve the numerical integration problem based on statistical sampling, the Monte Carlo method serves as the major approach in the constant temperature simulation. Following the Metropolis-Hasting algorithm\(^2\), in each step...
of the Monte Carlo sampling in canonical ensemble, a configuration move $X \rightarrow X'$ is made with the acceptance ratio

$$\text{acc}(X \rightarrow X') = \min \left\{ 1, \frac{P(X'; \beta)}{P(X; \beta)} \right\} = \min \{1, e^{-\beta \Delta E} \} ,$$

**Equation 2-1 Acceptance ratio in Metropolis algorithm**

where $\Delta E \equiv E(X') - E(X)$ denotes the potential energy difference between two configurations and $\beta$ is the reciprocal temperature. Therefore, this move is always accepted for $E(X') < E(X)$. The acceptance ratio satisfies the detailed balance, and the whole configurational space is expected to be sampled following the Boltzmann distribution given enough sampling time.

### 2.1.2. Molecular Dynamics

Another way to explore the configuration space is molecular dynamics, which is widely used in molecular systems. Different from Monte Carlo method which generate a new configuration by a random move, Molecular dynamics uses the classical Newton's equation to update the coordinates. Not only molecular dynamics can sample configurations in canonical ensembles, it allows one to observe the evolution of system in a physical timescale as well.

In molecular dynamics of an isolated system (microcanonical ensemble), the Newton's equation of the $i^{th}$ particle follows
\[
\begin{align*}
\dot{x}_i &= \dot{v}_i \\
\dot{v}_i &= \frac{\dot{F}_i}{m_i}
\end{align*}
\]

**Equation 2-2 Newton’s equation guiding molecular dynamics**

in which \(\vec{x}_i\) represents the coordinate, \(\vec{v}_i\) represents the velocity, \(m_i\) is the mass of this particle and \(\vec{F}_i = -\partial_{\vec{x}_i} U(\vec{x}_0, ..., \vec{x}_N)\) is the molecular force on this particle where \(U\) denotes the potential energy function and \(N\) is the particle number.

There are several approaches to integrate the Newton equation numerically. One of the most commonly used integrator is called the leap-frog, which updates the coordinates and velocities in the following way:

\[
\begin{align*}
\left\{ \begin{array}{l}
\quad x(t + \Delta t) = x(t) + v(t + \frac{1}{2} \Delta t) \Delta t \\
\quad v(t + \frac{1}{2} \Delta t) = v(t - \frac{1}{2} \Delta t) + \frac{F(t)}{m} \Delta t
\end{array} \right.
\]

**Equation 2-3 Leap-frog integrator in molecular dynamics**

where \(\Delta t\) is the time step and should be kept constant during the simulation. Although it is a low order integration method, the symplectic nature of leap-frog integrator guarantees the energy conservation even for long dynamics and generates time-reversible trajectories.

The equation above only shows how to simulate an isolated system in molecular dynamics. When dealing with the constant temperature simulation, the
concept of thermostat need to be introduced to provide a correct sampling in canonical ensemble, in which the system is imagined to have thermal contact with an infinite-large heat reservoir to keep the temperature fixed. There are several thermostats for constant temperature molecular dynamics each with different advantages, such as Langevin thermostat\(^3\), Nosé-Hoover thermostat\(^4\) and velocity-rescaling thermostat\(^5\). We only summarize the velocity-rescaling thermostat in the following paragraph, which is used in all of the molecular dynamics simulations in this thesis.

In the velocity-rescaling thermostat\(^5\), a four-step evolution of the system is used instead of the one-step evolution in an isolated system (microcanonical ensemble) simulation. These four steps are as follows:

1. Update the system using Newton equation
2. Calculate the current kinetic energy based on the current velocities
3. Do a random walk in kinetic energy space following the stochastic equation

\[
dK = \left( \frac{\bar{K} - K}{\tau} \right) dt + 2\frac{\sqrt{K\bar{K}}}{\sqrt{N_f\tau}} dW,
\]

**Equation 2-4 Kinetic energy random walk in velocity-rescaling thermostat**

where \(dW\) represents the Weiner process. Parameters \(\bar{K}, N_f, \tau\) represent the average kinetic energy in canonical ensemble, the degrees of freedom and a timescale constant of thermostat, respectively.
Multiply all the velocities by a factor $\alpha$ according to the updated kinetic energy, in which $\alpha = \sqrt{\frac{K}{\bar{K}}}$. The velocity recalling method allows a delay of the kinetic energy deviations from the desired value and thus generates a correct kinetic energy distribution in canonical ensemble.

**2.1.3. Limitations of NVT Simulation**

No matter Monte Carlo method or molecular dynamics simulation, a common issue of constant temperature simulation is the broken ergodicity, which can be caused by the rugged landscape of the energy function in the configurational space. For landscapes with various minima, the system is easy to get trapped in a local minimum. In Monte Carlo method, as the probability to make a move that increases the potential energy is smaller than 1, it may take a long time to jump out of a local basin if the energy difference between the bottom and the edge of this basin is quite large. In molecular dynamics simulation, as the fluctuation of kinetic energy is limited, the system can be trapped in the local minimum for a long time if the kinetic energy needed to jump out of the current basin is much larger than the average kinetic energy in the current canonical ensemble.

Although broken ergodicity does not actually exist for infinite time simulation, it can be a major obstacle of sampling the configurational space thoroughly given finite simulating time. What makes it worse is that the energy difference between the bottom and edge, or the “depth”, of energy basins often
increases with the system size, which makes it even more difficult for constant
temperature simulation to make effective sampling in large, complex systems with
rugged landscape.

2.2. Generalized Ensemble

2.2.1. Introduction

The idea of generalized ensemble was proposed a few decades ago\textsuperscript{7-9, 10, 11} to
overcome the broken ergodicity issue occurred in the \((N,V,T)\) simulation mentioned
above. As its name implies, a generalized ensemble uses a more general form of
probability function \(W(X, \{\lambda\})\), in which \(X\) represents the configuration and \(\{\lambda\}\)
represents the parameter set, instead of the Boltzmann distribution in canonical
ensemble. As energy and temperature are usually crucial in the simulation, the
forms of generalized ensemble are often related to these two quantities, i.e.
\(W(E(X); \beta)\) where \(E(X)\) is the potential energy of current configuration and \(\beta\) is
the reciprocal temperature. Generalized ensembles use various techniques to
accelerate the barrier-crossing speed and allow the system to explore the
configurational space efficiently for limited-length simulations.

Among all the generalized ensembles, there are three kinds of generalized
ensemble widely used in the field of molecular simulation, which are called multi-
canonical ensemble, simulated tempering and parallel tempering, respectively. We
will briefly summarize the main points of them in the next sections of this chapter.
2.2.2. Multi-canonical Ensemble

Multi-canonical ensemble\textsuperscript{10,12} is a technique that can sample various local minima in rugged landscapes effectively. The multi-canonical ensemble, or flat energy histogram method, uses a probability function \( W(X) \propto \Omega^{-1}(E(X)) = e^{-\beta T S(E(X))} \), in which \( \Omega(E(X)) \) is the density of states, \( T \) is the temperature and \( S(E(X)) \) is the microcanonical entropy. This mathematical form provides a constant probability distribution in energy space and generates the flat energy histogram correspondingly. In practice, the density of states \( \Omega \) is often unknown in priori and need to be estimated and update along with the simulation. Multi-canonical ensemble has already been shown to be very effective in simulating many complex systems such as the spin glass\textsuperscript{10,13}, simple liquid systems\textsuperscript{14}, small peptides\textsuperscript{15} and structure-based proteins\textsuperscript{16}.

2.2.3. Simulated Tempering

Another approach to sample the high-dimensional configurational space is to modulate the system temperature to avoid trapped in local minima. This approach is called simulated tempering\textsuperscript{11} and it can increase or decrease the temperature adaptively. The probability function in conventional simulated tempering has a form of \( W(T) = \text{const} \), which means a flat temperature histogram can be generated. In practice, the temperature is defined in a large range with \( m \) discrete values \( T_i \ (i = 1, \ldots, m) \) such that \( W(T_1) = W(T_2) = \cdots = W(T_m) \). In this case, the probability distribution of configuration \( X \) can be written as
\[ P(X) = \frac{1}{m} \sum_i e^{-\beta_i E(X)} Z(\beta_i). \]

**Equation 2-5 Configuraional probability distribution in simulated tempering**

Therefore, this distribution can be interpreted as a sum of canonical ensemble at different temperatures with equal weights.

During the simulation, an attempt move of temperature is made after several moves of configuration. To satisfy the detailed balance, the acceptance ratio of this attempt is

\[ acc(\beta_i \rightarrow \beta_j) = \min \left\{ 1, \frac{P(X; \beta_j)}{P(X; \beta_i)} \right\} = \min \left\{ 1, e^{-(\beta_j - \beta_i)E} \frac{Z(\beta_i)}{Z(\beta_j)} \right\} \]

**Equation 2-6 Acceptance ratio for temperature move in simulated tempering**

We can also write the acceptance ratio above as

\[ acc = \min \{ 1, \frac{P(X; \beta_i)\Omega(E(X))}{P(X; \beta_i)\Omega(E(X))} \} = \min \{ 1, \frac{P(E; \beta_i)}{P(E; \beta_i)} \}. \]

It can be seen that the successful exchange rate is highly dependent on the overlap area of the energy distribution of two adjacent canonical ensemble distributions.

Different from the multi-canonical ensemble, simulated tempering does not generate a flat energy histogram. However, an introduction of weight \( \{ w_i \} \) on each temperature \( T_i \) would also make the energy distribution quasi-flat\(^{17} \). Consider the height \( h_i \) of Boltzmann distribution at temperature \( T_i \), which is inversely
proportional to the width of this distribution $\sqrt{\langle (\Delta E)^2 \rangle_i}$. We can thus add a weight $w_i$ on the $i^{th}$ distribution which equals to $\sqrt{\langle (\Delta E)^2 \rangle_i} = T \sqrt{C_V}$ to make each distribution have the same height. Therefore, a quasi-flat energy histogram can also be generated from simulated tempering.

To satisfy a sufficient overlap of two adjacent energy distributions, the distance between the two peaks in each distribution, i.e. $\langle E \rangle_j - \langle E \rangle_i \propto C_V (T_j - T_i)$, needs to be comparable with the fluctuations in each distribution, i.e. $\sqrt{\langle (\Delta E)^2 \rangle_j}$ or $\sqrt{\langle (\Delta E)^2 \rangle_i}$. This leads to a requirement of the temperature difference $\Delta T$ between adjacent temperature values comparable with $\sqrt{\langle (\Delta E)^2 \rangle_j} / C_V \propto T / \sqrt{C_V}$. On the other hand, as the heat capacity $C_V$ increase with the system size, i.e. $C_V \propto N$ where $N$ indicates the system size, the required $\Delta T$ decrease with $N$ as $\Delta T \propto 1 / \sqrt{N}$. Therefore, an increasing number of $m \propto \sqrt{N}$ is needed with the increase of the system size.

Simulated tempering has also been applied to many applications such as spin model systems\textsuperscript{18}, molecular simulation of polypeptides\textsuperscript{19} and many other global optimization problems\textsuperscript{20}. It is worth mentioning that in small peptide simulation, the performance of multi-canonical ensemble and simulated tempering was shown to be similar\textsuperscript{21}. However, the latter method does not need to estimate $\Omega(E)$ or $S(E)$ of the system precisely\textsuperscript{17, 22, 23}, which makes it easier to be applied to in larger systems, such as the protein folding simulation in explicit solvent\textsuperscript{23, 24}. Simulated tempering is particularly useful in situation that the energy change upon folding polypeptide is smaller than the total energy fluctuation of the entire system when a large water box is present.
2.2.4. Parallel Tempering

Parallel tempering\textsuperscript{7,25}, or replica exchange, provides another way of searching different regions and minima in the rugged landscape. In the parallel tempering scheme, multiple copies are used and each copy samples in canonical ensembles at different temperatures. The joint distribution of configuration $X$ is

$$P(\{X_i\}) = \prod_i e^{-\beta_i E(X_i)} / Z(\beta_i),$$

\textbf{Equation 2-7 Configuraional probability distribution in parallel tempering}

and this will definitely generate a flat temperature distribution, i.e. $W(T_1) = W(T_2) = \cdots = W(T_m)$ where $m$ is the number of copies.

In parallel tempering, an attempt of exchanging the temperatures between adjacent copies is made at certain time intervals. Following the condition of detailed balance, the acceptance ratio can be written as

$$\text{acc}(X_i \beta_i X_j \beta_j \rightarrow X_i \beta_j X_j \beta_i) = \min \left\{ 1, \frac{P(X_i; \beta_j)P(X_j; \beta_i)}{P(X_i; \beta_i)P(X_j; \beta_j)} \right\} = \min \left\{ 1, e^{\Delta \beta \Delta E} \right\},$$

\textbf{Equation 2-8 Acceptance ratio for replica exchange in parallel tempering}

where $\Delta \beta \equiv \beta_i - \beta_j$ and $\Delta E \equiv E(X_i) - E(X_j)$. In practice, $m$ processes or threads are running simultaneously, with each process or thread represents a certain temperature. If the exchange attempt is accepted, the temperature in each process
or thread remains unchanged while the configuration information, e.g. coordinates and velocities, is exchanged between involved copies.

There are two biggest advantages in parallel tempering. The first one is to enable each copy to sample a region which is totally uncorrelated to the previous one, which is particularly useful in the application of phase transition problems in which the correlation time near the critical temperature can be very long\(^8\). Another reason is that the introduction of multiple copies brings more data, which can be combined to make a precise estimation of physical properties using reweighting techniques such as weighted histogram analysis method (WHAM)\(^26\). In addition, it is easy to implement compared to simulated tempering in that the partition function \(Z(\beta)\) is not required to compute the acceptance ratio. On the other hand, the limitation in conventional parallel tempering is also obvious. Similar to the conventional simulated tempering, the number of copies required to keep a reasonable acceptance ratio increases with the system size\(^27\), i.e. \(m \propto \sqrt{N}\). This requirement makes the parallel tempering method consume significant computing resources for large systems. Although progresses\(^28\) have been made to increase the exchange probability of parallel tempering in recent years, this issue remains to be a major obstacle in the application of parallel tempering in large systems.
3.1. Continuous Simulated Tempering (CST)

On the basis of the fact that the increasing number of discrete temperature is needed for large systems in conventional simulated tempering, CST was proposed to set the temperate in continuous values. In CST, the reciprocal temperature $\beta$ is defined in a large range $(\beta_{\text{min}}, \beta_{\text{max}})$ which covers all the interested temperatures for the targeted system. The probability distribution of $\beta$ and the configuration $X$ is defined as

$$P(X, \beta) = CZ^{-1}(\beta) e^{-\beta E(X)} \omega(\beta),$$

Equation 3-1 Probability distribution of configuration in CST
in which $E(X)$ is the potential energy of configuration $X$, $Z(\beta)$ is the canonical partition function $\omega(\beta)$ is the temperature probability distribution with normalization factor $C$. $\omega(\beta)$ can also be interpreted as the weight of the canonical ensemble at the temperature $\beta$. It was discussed in the last chapter that a chosen of $\omega(\beta) \propto \sqrt{C_V(\beta)}/\beta$ can generate a quasi-flat energy histogram and keep a sufficient overlap of energy histograms at adjacent temperatures in canonical ensemble. For example, for the systems with a roughly constant heat capacity (such as ideal gas or Lennard-Jones system), the optimized $\omega(\beta)$ would be $\beta^{-1}$ and for other systems like proteins in which heat capacity was observed as $\propto \beta$, a suggested $\omega(\beta)$ would be $\beta^{-0.5}$. Generally speaking, $\omega(\beta)$ can be represented as $\beta^{-\gamma}$ if $\sqrt{C_V(\beta)} \propto \beta^{1-\gamma}$.

To generated the desired $P(X, \beta)$, a temperature random walk scheme is applied instead of the acceptance ratio scheme in the conventional simulated tempering. In this case, a general form of 1-dimensional Fokker-Planck equation follows

$$\frac{\partial \rho(T)}{\partial t} = -\frac{\partial}{\partial T} [\mu(T) \rho(T)] + \frac{\partial^2}{\partial T^2} [D(T) \rho(T)] ,$$

Equation 3-2 General form of one-dimensional Fokker-Planck equation

in which $\rho(T)$ is the instant temperature probability distribution, $D(T)$ is a diffusion term and $\mu(T)$ represents a drift force in temperature space with the mathematical form of $\partial_T[D(T)P(X,T)]/P(X,T) = D(T)\partial_T \ln P(X,T) + \partial_T D(T)$. Based on the
equivalence of Brownian motion and 1-dimensional Fokker-Planck equation in this form, a temperature space random walk can be written as

\[ dT = \left[ D(T) \frac{\partial \ln P(X,T)}{\partial T} + \frac{\partial D(T)}{\partial T} \right] dt + \sqrt{2D(T)} dW, \]

**Equation 3-3 Random walk in temperature space**

where \( dW \) is the Weiner process. It can be easily shown that the stationary distribution of this F-P equation is \( P(X,T) \). This Brownian equation is valid for any form of \( D(T) \), while \( D(T) = T^2 \) was chosen\(^{23}\) to encourage the diffusion at high temperatures.

Considering that \( P(X,T) = P(X,\beta)|\partial_T \beta| = P(X,\beta)/k_B T^2 \), the same Brownian equation but in the \( \beta \)-representation can be derived as

\[ d(\frac{1}{\beta}) = -\frac{\partial \ln P(X,\beta)}{\partial \beta} dt + \frac{\sqrt{2}}{\beta} dW = \left( E - \bar{E}(\beta) - \frac{\partial \ln \omega(\beta)}{\partial \beta} \right) dt + \frac{\sqrt{2}}{\beta} dW, \]

**Equation 3-4 Random walk in temperature space**

in reciprocal temperature representation

where \( \bar{E}(\beta) \equiv -\beta \ln Z(\beta) \) denotes the average potential energy at current temperature in canonical ensemble and \( E \) is the potential energy of current configuration which is evaluated during the simulation.

CST has already been shown to be very successful in simulating large, complex systems\(^{23}\) in that it provides a high barrier-crossing efficiency while
overcoming the issue of the decreasing acceptance ratio with the increase of system size for fixed-number of discrete temperatures in conventional simulating tempering. However, there are systems with special interest in some important temperature ranges, such as near room temperature for protein folding simulation or near critical temperature in phase transition simulation. These systems often have longer relaxation time\textsuperscript{29} (such as folding a long polypeptide) or correlation time\textsuperscript{8} (such as phase transition). In this situation, although CST could sample the whole temperature range, the limited residing time in the important temperature ranges affects the searching efficiency in this temperature range. For example, in the application of protein folding, the interested temperature range is quite low (\textasciitilde300K) while the temperatures needed to cross barriers are much higher (> 500K). In this situation, there exists a dilemma between barrier-crossing efficiency and the demand of importance sampling around the low temperature range. To solve this dilemma and enhance the sampling near the important temperature range, we propose a parallel continuous simulated tempering (PCST) with a revised generalized ensemble form, which will be discussed in the next section.

3.2. Parallel Continuous Simulated Tempering (PCST)

As its name implies that PCST employs multiple-copies which collaborate with each other to enhance the sampling near important temperature ranges. In PCST, the probability distribution of configuration $X$ and reciprocal temperature $\beta$ in the $i^{\text{th}}$ copy has a mathematical form of
\[ P(X, \beta) = CZ^{-1}(\beta) e^{-\beta E(X)} \beta^{-\gamma} f_i(\beta). \]

**Equation 3-5 Probability distribution of configuration in PCST**

The function \( f_i(\beta) \) can be designed to either increase or decrease the weight of canonical ensemble at temperature \( \beta \) and the sampling time on certain temperature ranges. For example, a Gaussian function \( e^{-(\beta - \beta_i^0)^2 / 2\sigma_i^2} \) can focus the importance sampling around the interested temperature \( \beta_i \). Following the same derivation as the one in the CST method, the random walk in temperature space can be written as

\[
d \left( \frac{1}{\beta} \right) = \left( E - \bar{E}(\beta) - \frac{\partial \ln \omega(\beta)}{\partial \beta} - \frac{\partial \ln f_i(\beta)}{\partial \beta} \right) dt + \frac{\sqrt{2}}{\beta} dW,
\]

**Equation 3-6 Random walk in temperature space in PCST**

with the Langevin equation

\[
\frac{d(1/\beta)}{dt} = E - \bar{E}(\beta) - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2} + \frac{\sqrt{2}}{\beta} \xi,
\]

**Equation 3-7 Corresponding Langevin equation of Equation 3-6**

where \( \gamma \) is either 0.5 or 1 according to the specific system and \( \xi \) is the standard Gaussian noise.

In order to facilitate barrier crossing at high temperature and sampling at lower temperature thoroughly, we introduce a parameter exchange protocol...
between high temperature copies and low temperature copies. In this protocol, the acceptance ratio for copy exchange between the \(i^{th}\) and \(j^{th}\) copy is defined as

\[
\text{Acc}(X_i\beta_i\beta_i^0\sigma_i, X_j\beta_j\beta_j^0\sigma_j \rightarrow X_i\beta_j\beta_j^0\sigma_i, X_j\beta_i\beta_i^0\sigma_j) \\
= \min\left\{1, \frac{P(X_j, \beta_j; \beta_j^0, \sigma_j)P(X_i, \beta_i; \beta_i^0, \sigma_i)}{P(X_i, \beta_i; \beta_i^0, \sigma_i)P(X_j, \beta_j; \beta_j^0, \sigma_j)}\right\} = \min\left\{1, e^{-(\Delta\beta\Delta q_0 + \Delta\beta^2\Delta q_1)}\right\}.
\]

**Equation 3-8 Acceptance ratio for parameter exchange in PCST**

where \(\Delta\beta \equiv \beta_i - \beta_j\), \(\Delta\beta^2 \equiv \beta_i^2 - \beta_j^2\), \(\Delta q_0 \equiv \frac{\beta_i^0}{\sigma_i^2} - \frac{\beta_j^0}{\sigma_j^2}\) and \(\Delta q_1 \equiv -\frac{1}{2\sigma_i^2} + \frac{1}{2\sigma_j^2}\).

Specially, if the variances of different distributions are the same, i.e. \(\sigma_i = \sigma_j\), then the acceptance ratio becomes

\[
\text{Acc}(X_i\beta_i\beta_i^0, X_j\beta_j\beta_j^0 \rightarrow X_j\beta_i\beta_i^0, X_i\beta_j\beta_j^0) \\
= \min\left\{1, \frac{P(X_j, \beta_j; \beta_j^0)P(X_i, \beta_i; \beta_i^0)}{P(X_i, \beta_i; \beta_i^0)P(X_j, \beta_j; \beta_j^0)}\right\} = \min\left\{1, e^{-\Delta\beta^0/\sigma^2}\right\},
\]

**Equation 3-9 Acceptance ratio in PCST for same variances in involved distribution**

where \(\Delta\beta = \beta_i - \beta_j\) and \(\Delta\beta^0 = \beta_i^0 - \beta_j^0\).

It is easy to show that this protocol satisfies the detailed balance. An important feature of the parameter exchange protocol is that the acceptance ratio is only related to the parameters in generalized ensemble and the current temperature, whereas it does not have an explicit dependence on the current
potential energy $E$. Therefore, compared to the conventional replica-exchange method, PCST does not suffer from a decrease of successful exchange rate with the increase of system size. Moreover, the acceptance ratio is user-adjustable via changing the parameters in generalized ensemble. It is noted that at the limit $\sigma \to 0$, the generalized ensemble is converted into isolated canonical ensembles at temperature $\beta_i^0$ for the $i^{th}$ copy. In this case, if the temperature gap $\Delta \beta^0$ is relatively large, then the exchange ratio will decrease to zero.

### 3.3. Estimating Physical Properties in Simulated Tempering

#### 3.3.1. Introduction

As shown in the Equation 3-6 that among all the quantities needed to guide the Langevin equation, the only one unknown in priori is the average energy $\bar{E}(\beta)$. Meanwhile, it is essential to know $\bar{E}(\beta)$ to estimate other quantities, such as heat capacity $C_V = \partial_T \bar{E}$ and the logarithm of partition function $\ln Z(\beta) = \ln Z(\beta_{\min}) - \int \bar{E}(\beta) d\beta$ which is related to the free energy. Therefore, estimating $\bar{E}(\beta)$ becomes one of the key parts in CST/PCST.

During the simulated tempering simulation, many small bins $(\beta_i, \beta_{i+1})$ are created in a large temperature range $(\beta_{\min}, \beta_{\max})$. If the current temperature falls into the $i^{th}$ bin, then the current potential energy is collected to calculate the average energy in this bin $E_i$, which is defined as $\frac{1}{\beta_{i+1} - \beta_i} \int_{\beta_i}^{\beta_{i+1}} E(\beta) d\beta$. Since the bin size is very small, we assume that the average potential energy in this bin is
constant, which is already shown to be valid even for finite-size phase transition problems. In this case, the Langevin equation can be approximately written as

\[ \frac{d(1/\beta)}{dt} = E - \bar{E}_i - \frac{\gamma}{\beta} + \frac{\beta - \beta_i^0}{\sigma_i^2} + \frac{\sqrt{2}}{\beta} \xi. \]

**Equation 3-10** Langevin equation after introducing multiple bins for collection of data

Since the energy data in a single bin may not be reliable, especially in the early stage of simulation when the system has not equilibrated properly, several advanced techniques such as adaptive averaging\(^{17}\) and multiple-bin estimator based on integral identity\(^{23}\) (MEII) are used. These techniques in enhancing the accuracy of estimating \(\bar{E}_i\) were clearly shown in previous work. We briefly summarize the main points of adaptive averaging and MEII in the next sections.

### 3.3.2. Adaptive Averaging Techniques

Adaptive averaging\(^ {17}\) reduces the negative effect of the data generated in the early stages of simulation on estimating \(\bar{E}_i\). Suppose the \(k\)th sample of potential energy in the \(i\)th small bin is \(E_i^{(k)}\) and its weight is \(w_i^{(k)}\). If we define \(S_d^{(n)} = \sum_{k=1}^{n} w_i^{(k)} E_i^{(k)}\), then the weighted mean and variance of potential energy and its square in this bin can be written as

\[
\langle E \rangle_i^{(n)} = \frac{S_1^{(n)}}{S_0^{(n)}}
\]

\[
\langle (\Delta E)^2 \rangle_i^{(n)} = \langle E^2 \rangle_i^{(n)} - \langle E \rangle_i^{2(n)} = \left( S_2^{(n)} - \frac{S_1^{(n)} S_0^{(n)}}{S_0^{(n)}} \right) \equiv \sigma_i^{(n)} / S_0^{(n)}.
\]
**Equation 3-11 Weighted mean and variance in each small bin**

The adaptive averaging scheme uses \( w_i^{(k)} = \frac{1}{1 - \frac{C}{k}} \), where \( C \) is a constant which is smaller than 1 and gradually reduces \( w^{(k)} \) from \( 1/(1 - C) \) to 1 as \( k \to \infty \).

In practice, \( S_0^{(n)}, S_1^{(n)} \) and \( \sigma^{(n)} \) are calculated incrementally as

\[
S_0^{(n)} = S_0^{(n-1)} + w_i^{(n)} \\
S_1^{(n)} = S_1^{(n-1)} + w_i^{(n)} E_i^{(k)} \\
\sigma^{(n)} = \sigma^{(n-1)} + w_i^{(n)} \left( E_i^{(k)} - S_1^{(n-1)}/S_0^{(n-1)} \right) \left( E_i^{(k)} - S_1^{(n)}/S_0^{(n)} \right).
\]

**Equation 3-12 Incremental calculation of weighted mean and variance**

### 3.3.3. Multi-bin Estimator based on Integral Identities (MEII)

Although the negative effect of early simulation data can be eliminated by the adaptive averaging, the data collected in each single bin is still far from reliable. This is mainly due to the fact that as the large number of bins is quite reduces the number of samples in each small bin in a finite time simulation. In this situation, the convergence speed of \( \langle E \rangle_i^{(n)} \) and \( \langle \Delta E^2 \rangle_i^{(n)} \) are quite low which reduces the accuracy of estimated \( \bar{E}_i \). To overcome this limitation, a multiple-bin estimator based on integral identity\(^\text{23} \) (MEII) was proposed.

The main idea of MEII is to borrow the statistics in neighboring bins to provide a much more accurate estimation of \( \bar{E}_i \) (we will call it \( \bar{E}_i \) in the following
In canonical ensemble, the heat capacity has a direct relationship with the fluctuation of potential energy at the same temperature

\[ C_V(T) \equiv \left( \frac{\partial \overline{E}}{\partial T} \right)_{N,V} = \frac{1}{k_B T^2} \langle \Delta E \rangle^2 . \]

**Equation 3-13 Relationship between heat capacity and energy fluctuation in canonical ensemble**

We can also write the equation above as \( \partial_\beta \overline{E} = -\langle \Delta E \rangle^2 \), which brings the integral identity for any continuous function \( \phi(\beta) \) with zero values at \( \beta_+ \) and \( \beta_- \) as

\[
\int_{\beta_-}^{\beta_+} \phi'(\beta) \overline{E}(\beta) d\beta = \phi(\beta) \overline{E}(\beta) \bigg|_{\beta_-}^{\beta_+} - \int_{\beta_-}^{\beta_+} \phi(\beta) \frac{\partial \overline{E}(\beta)}{\partial \beta} d\beta = \int_{\beta_-}^{\beta_+} \phi(\beta) \langle \Delta E \rangle^2 d\beta .
\]

**Equation 3-14 Integral identity for any continuous function**

Given the approximation that \( \overline{E}(\beta) = \overline{E}_i \) and \( \langle \Delta E \rangle^2 = \langle \Delta E \rangle^2_i \) in each small bin, the integral identity above can be written in the vector form of

\[ \Phi^T E = \Phi^T F \]

**Equation 3-15 Integral identity in Equation 3-14 in vector form**

where \( \Phi_i \equiv \int_{\beta_i}^{\beta_{i+1}} \phi'(\beta) d\beta ; \quad \Phi_i \equiv \int_{\beta_i}^{\beta_{i+1}} \phi(\beta) d\beta ; \quad E_i \equiv \overline{E}_i ; \quad F_i \equiv \langle \overline{E} \rangle^2_i \). Thus, the energy average in the \( j^{th} \) bin \( \overline{E}_j \) is exactly equal to
\[
\bar{E}_j = -\Phi_i E_i + \phi_i F_i + \bar{E}_j = \psi_i E_i + \phi_i F_i,
\]

where \(\psi_i \equiv -\Phi_i + \delta_{ij}\).

In real simulation, while \(\bar{E}_i\) and \((\Delta E)^2\) are substituted by sample average \(\langle E \rangle_i\) and \((\langle \Delta E \rangle)^2_i\) on the right hand side of the equation above, the form of \(\phi(\beta)\) can be selected to keep \(\phi^T F\) always equal to zero to reduce the sampling error. In this case, an unbiased estimation of any interested \(\bar{E}_j\) becomes

\[
\bar{E}_j = \psi^T E.
\]

**Equation 3-16 Unbiased estimation of average energy in each small bin**

In the previous work\(^{23}\), the form of \(\phi(\beta)\) was given as

\[
\phi(\beta) = \begin{cases} 
  a_- \frac{\beta - \beta_-}{\beta_{j+1} - \beta_-}, & \beta \in (\beta_{-}, \beta_{j}) \\
  a_- \frac{\beta - \beta_-}{\beta_{j+1} - \beta_-} - \frac{\beta - \beta_j}{\beta_{j+1} - \beta_j}, & \beta \in (\beta_{j}, \beta_{j+1}) \\
  a_+ \frac{\beta - \beta_+}{\beta_+ - \beta_{j+1}}, & \beta \in (\beta_{j+1}, \beta_{+})
\end{cases}
\]

**Equation 3-17 Mathematical form of \(\phi(\beta)\)**
so \( \psi_i = \begin{cases} \frac{a_-}{\beta_{j+1} - \beta_-}, & i \leq j \\ \frac{a_+}{\beta_+ - \beta_{j+1}}, & i > j \end{cases} \)

and \( \phi_i = \begin{cases} a_- \frac{\beta_i - \beta_-}{\beta_{j+1} - \beta_-} \Delta \beta_i, & i < j \\ a_+ \frac{\beta_j - \beta_-}{\beta_{j+1} - \beta_-} \Delta \beta_i - \frac{1}{2} \Delta \beta_j, & i = j \\ a_+ \frac{\beta_i - \beta_j}{\beta_+ - \beta_{j+1}} \Delta \beta_i, & i > j \end{cases} \)

**Equation 3-18 Specific value of \( \phi \) and \( \psi \) in each small bin**

where \( \bar{\beta}_i \equiv \frac{\beta_i + \beta_{i+1}}{2} \) and \( \Delta \beta_i \equiv \beta_{i+1} - \beta_i \). Usually \( \beta_i \), which is the bin size of the \( i^{th} \) bin, is a constant and does not change with the index \( i \). In this scheme, two parameters \( a_- \) and \( a_+ \) are both positive and their sum is 1. The parameters \( a_- \) and \( a_+ \) also represent the “weight” of the sum of \( \psi_i E_i \) on the left and right side of the \( j^{th} \) bin, and they are determined by the equation \( \phi_i F_i = 0 \). In practice, the value of \( \phi_i \) and \( \psi_i \) do not change with time and only need to be calculated once. Moreover, and the window \( (\beta_-, \beta_+) \) is chosen so that the interested bin are approximately in the middle of the window so the solutions of \( a_- \) and \( a_+ \) are easily to get positive values.

In summary, the implementation of MEII can be divided into two major steps:

1. Given the current sample variance \( F_i = \langle (\Delta E)^2 \rangle_i \) of every small bins in the window, get \( a_- \) and \( a_+ \) using the equation \( \phi_i F_i = 0 \) given the form of \( \phi_i \) in Equation 3-18;

2. Given the current sample mean of \( E_i = \langle E \rangle_i \) of every small bins in the window and parameters \( \{a_-, a_+\} \), get \( \bar{E}_j \) using the equation \( \bar{E}_j = \psi_i E_i \) given the form of \( \phi_i \) in Equation 3-18.
In this way, $\bar{E}_j$ can serve as an unbiased estimation of the exact average potential energy $\bar{E}_j$ in the $j^{th}$ bin, and it will converge to $\bar{E}_j$ asymptotically with long time simulation.

### 3.4. Simulation Protocols of PCST

There are three time intervals in parallel generalized ensemble simulation: $dt_{CE}$, the timestep in canonical ensemble at a fixed temperature, i.e. the simulation timestep; $dt_{\text{walk}}$ the timestep for integrating the Langevin equation guiding the temperature random walk; and $dt_{\text{ex}}$ the time interval to attempt an exchange of parameters between different copies. The detailed simulation protocols are outlined below.

1. For each copy, run a short trajectory in canonical ensemble with the timestep $dt_{CE}$, and record the potential energy (if available).
2. For every time interval of $dt_{\text{walk}}$, (a) if current temperature $\beta$ falls into the $i^{th}$ bin, then estimate the average energy $\bar{E}_i$ using MEII; (b) update temperature based on the Langevin equation (Equation 3-10).
3. For every time interval of $dt_{\text{ex}}$, (a) collect the statistics from every copy and combine them; (b) based on the current temperature $\beta$ and parameters $\{\beta^0, \sigma\}$ in selected pairs of copy and calculate the acceptance ratio using Equation 3-8; (c) if accepted, exchange the states (coordinates $x$, velocities $v$, temperature $\beta$ and thermostat-related parameters, etc) of two copies while keeping generalized ensemble parameters unchanged.
3.5. Applications

3.5.1. Phase Transition Simulation in 2-dimentional Ising Model

We first test our method on a 32 × 32 Ising model to show how our generalized ensemble works in phase transition problems. We have set the temperature range (0.25, 0.65) which covered the critical temperature of the phase transition. The bin size \( \delta \beta \) equals to 0.0002. The parameter \( \gamma \) in Equation 3-5 was set as zero, which corresponds to flat-\( \beta \) histogram. For each temperature bin, MEII was implemented with the window size of 201 bins. The Langevin equation was integrated after every 100 Monte Carlo moves, with an integrating step \( \Delta t = 2 \times 10^{-5} \).

We performed the simulation with length of \( 1.024 \times 10^9 \) steps which corresponds to \( 10^6 \) flips per site. Two copies are applied and the parameter \( \{ \beta_0, \sigma \} \) in each copy was set as \( \{0.5, 0.1\} \) and \( \{0.4, 0.1\} \). The heat capacity compared with analytical results is shown in Figure 1. It can be seen that the errors in heat capacity are relatively small compared to the analytical results. The temperature histograms of each copy are shown in Figure 2. It is shown that the peak positions and widths of temperature histograms follow the desired Gaussian distribution very well. This indicates that the PCST method can get an accurate estimation of physical properties for this phase transition problem.
Figure 1 Calculated heat capacity using PCST and the analytical results. The absolute error of calculated one is shown in the embedded figure.
Figure 2 Generated temperature distribution in the low temperature copy and high temperature copy

3.5.2. Lennard-Jones Fluid

To demonstrate the convergence of physical properties with time, we tested different methods on the Lennard-Jones fluid with 1000 particles. There are three methods were used in total: the previous CST method, PCST method with and without parameter exchange. In the reduced units, the thermostat temperature and density were set as 0.9 and 0.78, respectively. Following the convention of Rowley, Nicholson and Parsonage model in Argon\textsuperscript{30}, the thermostat temperature and the length of system cube were 107.82 K and 3.699 nm, respectively. The force
parameters \( \{C_6, C_{12}\} \) in Lennard-Jones potential were \( \{6.209 \times 10^{-3}, 9.677 \times 10^{-6}\} \) kJ mol\(^{-1}\) nm\(^6\). A force-rescaling scheme\(^{17}\) was applied to sample at different temperature using the fixed-temperature thermostat. The temperature range for tempering was set as \( \beta = (0.5, 1.11) \). Two copied are used and the parameter \( \{\beta_0, \sigma\} \) in each copy was set as \( \{0.65, 0.3\} \) and \( \{0.95, 0.3\} \), respectively. Molecular dynamics simulation was implemented with an integration time step \( \Delta t = 0.0002 \). The total simulation length is \( 1 \times 10^9 \) steps and set the estimated average energy as reference. For each method, we calculated the absolute error \( \Delta E \) at different time from the reference one by averaging all the absolute errors in every involved bins in the targeted temperature range.

We show the speed of convergence of estimated average energy at higher temperature range (\( \beta = 0.5 \sim 0.7 \)) and lower temperature range (\( \beta = 1.0 \sim 1.11 \)) in Figure 2. At high temperature range, the CST method with high temperature bias ("high T" in Figure 2) shows the highest convergence speed, while the one with low temperature bias ("low T" in Figure 2) shows the highest convergence speed at low temperature range. For the original CST method (without any Gaussian temperature bias), the convergence at the high temperature range is similar to the one with high temperature bias while the convergence at the low temperature range is not as good as the one with low temperature bias. In PCST method, the convergence at low temperature range is enhanced, while the convergence at high temperature can still be kept similar to the CST method with high temperature bias. The result demonstrates that the exchange protocol in PCST method can accelerate the
convergence of physical properties at the whole temperature range, especially at the low temperature range compared with the other methods.

![Figure 3](image)

**Figure 3** The convergence of energy per atom in low temperature range and high temperature range (embedded) for every method.

### 3.5.3. Folding Small Proteins in All-atom Model

We tested our method in folding a 20 amino acid helical protein, tryptophan cage\(^{31}\) (pdb code 1L2Y, sequence NLYIQ WLKDG GPSSG RPPPS), which has already been studied by continuous ST method in our previous work\(^{23}\). It has an N-terminal helix, a short 3\(_{10}\)-helix and a C-terminal polyproline region. The polyproline region
is relatively flexible and does not form any secondary structure. For comparison, we also used the CST method for the folding study. We used AMBER99SB force field in all simulations.

We have implemented our method into a modified GROMACS 4.6.3 package. In all the simulations, the trp-cage protein was put in a cubic $46 \times 46 \times 46 \text{Å}^3$ box with 3143 TIP3P model of water molecules and one Cl$^-$ ion to keep the total charge zero. The cut-off distances of Lennard-Jones interaction, electrostatic interaction and neighbor list were 15 Å. Particle mesh Ewald (PME) method was used to calculate the electrostatic force, with the 1.19 Å grid spacing of Fourier transform. The parallel LINCS algorithm was used to constrain the hydrogen bonds in protein and the SETTLE algorithm was used for constraints in water molecules. The simulation was performed in an (N, V, T) ensemble with thermostat temperature 300K. Velocity rescaling scheme was applied on the thermostat. A force-rescaling scheme was applied to sample at different temperature using the fixed-temperature thermostat. The timestep for molecular dynamics integration was 0.002 ps. The timestep $dt_{\text{walk}}$ for integrating the Langevin equation was 0.04 ps, which was the same as the neighbor list refreshing interval. For our PCST method, the time interval $dt_{\text{ex}}$ of exchange attempt was 10 ns ($5 \times 10^6$ steps).

For both CST and PCST methods, we performed five independent trajectories with length of 2 μs for each method. The temperature range $(\beta_{\text{min}}, \beta_{\text{max}})$ was set as $(0.24, 0.41)$ kJ$^{-1}$mol and the length of small bin for data collection was
δβ = 0.0001. The window width for MEII was 200 times of the bin size, which was approximately 10% of the total temperature range. The parameter C in adaptive averaging\textsuperscript{17} was 0.1. Following the convention in the CST method\textsuperscript{23}, the factor γ in generalized ensemble was set as 1, which would make the system bias towards the high temperature\textsuperscript{23} and accelerate the barrier crossing. In our PCST method, this bias was removed and it was kept as 0.5. The parameters \{β_o, σ\} in Equation 3-5 were set as \{0.38, 0.05\} for the low temperature copy and \{0.27, 0.13\} for the high temperature copy (with the peaks of temperature distribution in 316K and 445K), respectively.

The folding results are indicated by alpha-carbon root mean square deviations (Cα-RMSD). The native state is marked by the blue line which represents 3 Å RMSD from the NMR structure\textsuperscript{31}. The results of five trajectories of CST method are shown in Figure 4. All five trajectories reach the native state for a few times. However, the length of time for the system to stay near the native state in each folding event is not very long. The average fraction of time to stay in the near native state of five trajectories was around 3%, with the smallest value 0.8% (traj 2) and the largest value 6% (traj 5). This may be partly due to high temperature bias was applied to satisfy the barrier-crossing efficiency, and it significantly reduces the time of sampling in the low temperature range. As we mentioned previously, to increase the low temperature sampling, it is far from enough to simply change the high temperature bias to low temperature bias, e.g. to change the generalized ensemble parameter γ from 1.0 to 0.5 or even smaller values, as the system may be trapped in energy basins.
Figure 4 C-alpha RMSD from the native state in five trajectories in CST simulation

In Figure 5, we show the C$_\alpha$-RMSD of five trajectories using PCST method. It can be seen that the occurrence of folding events was significantly increased comparing with that of CST method. Two copies are used in our simulation, one for high temperature and the other for low temperature. For the low temperature copy, the average fraction of time near the native state was 23.7%, with the smallest value 12% and the largest value 42%. For the high temperature copy, the average fraction was 12.8%, with the smallest value 7% and the largest value 20%. The results were achieved by the collaboration of two copies: the low temperature copy is mainly
responsible for searching in the folding temperature range while the high
temperature copy is mainly responsible for barrier crossing.

The relationship between RMSD and radius of gyration (Rg) of Ca atoms in
PCST method for both copies is shown in Figure 6. As the peaks of temperature
distribution in two copies are 445K and 316K respectively while the experimental
folding temperature is 315K,\(^{31, 38}\) we can approximately use the RMSD- Rg
relationship of two copies to observe the compaction of structure above or around
the folding temperature. It is clear that both simulation copies have substantial

Figure 5 C-alpha RMSD from the native state in five trajectories
in PCST simulation
populations at near native state. In Figure 6(a), the copy samples at high
temperature have a significant population in a region around 5Å RMSD from the
native state. In Figure 6(b), for the copy samples around the folding temperature,
the near native state region is the only favored region. It is worth mentioning, in
high temperature copy simulation, that the Rg of the population peak around 5Å
RMSD is similar to the native state, indicating that this peak corresponds to a state
that the hydrophobic collapse has already happened (molten globular state) and the
free energy difference between this state and the native state is mainly entropic. The
barrier between this state and the native state is very eminent. Such results indicate
that the generalized ensemble sampling methods are useful in exploring free energy
landscape and characterizing the folding transition states, even though the timescale
in this kind of simulations is not realistic.
The normalized RMSD distributions of two methods are shown in Figure 7. Compared to the CST method, the PCST method allows the system to sample near
the native state more thoroughly, so the population in the low RMSD range is
enhanced. Furthermore, in the distribution of the PCST method, two distinct peaks
were observed around 1Å and 2Å RMSD, respectively. This seems to suggest that
there are two substates near the native state. This phenomenon can also be easily
seen in Figure 5. We show the amplified parts of trajectory 3 and 5 (Figure 8) along
with the molecular structures of the two substates (Figure 9). Between the two
substates, the N-terminal helixes of the two structures are almost the same, and
differences between the two are mainly caused by the C-terminal polyproline
region. Our observation is consistent with the previous computational\textsuperscript{39,40} and
experimental\textsuperscript{40,41} studies about the existence of these two substates.
Figure 7 (a) The C-alpha RMSD distribution in PCST method. The dashed line and solid line denote the distribution in the high temperature copy and low temperature copy, respectively. (b) The C-alpha RMSD distribution in CST method.
Figure 8 Enlargement of parts of traj 3 and traj 5 in Figure 5
Figure 9 Molecular structure of two substates. The two structures are aligned with respect to the N-terminal helix (residues 1-10). Two important residues, Pro17 and Pro19, whose side chains show big difference between two substates, are explicitly drawn. Another residue Ser14, from which large difference of C-alpha trace between the two substates can be seen, is also indicated. The difference between the two substates is mainly in the C-terminal proline-rich loop.

To further characterize the nature of the two substates, we analyzed the $C_\alpha$-$C_\alpha$(helix) RMSD distribution for the PCST method (Figure 10), where $C_\alpha$(helix) represents the $C_\alpha$ atoms in the N-terminal helix formed by the first 10 residues. Results of both copies show that the two substates have almost the same $C_\alpha$(helix) RMSD, and the difference in the $C_\alpha$ RMSD is caused by the long polyproline in C-terminal region. We have found that there seems to be two distinct types of pathway from the collapsed state (the 5Å RMSD region mentioned before) to the native state,
one type is direct folding to the either substate of the native state (indicated by arrow I and II in Figure 10) and the other type is first folding the N-terminal helix to the native state followed by the folding of flexible polyproline region to the native state (indicated by arrow III in Figure 10). As the two substates were not observed in the CST method, the results of PCST method demonstrate that the existence of the low temperature copy provides more thorough sampling of the free energy landscape near the native state.
Figure 10 Normalized joint distribution of the RMSD of $\mathrm{C}_\alpha$ (helix) versus RMSD of $\mathrm{C}_\alpha$ atoms in
(a) high temperature copy and (b) low temperature copy.
In both subfigures, three pathways to the native state are labeled as I, II and III.
3.6. Discussions

In this chapter, we review the CST method and present a PCST method which enables copies of simulation to explore the configurational space collaboratively. A Gaussian distribution in temperature space with different parameters is added on each copy so that it can focus on searching the temperature range around the peak of Gaussian distribution. An exchange protocol of parameters is introduced to eliminate the broken ergodicity issue that might exist in the low temperature copies. The PCST has shown its effectiveness in simulating large, complex systems with rugged landscape and huge degrees of freedom and estimating physical properties in a precise way.

In the simulation of 2-dimensional Ising model and Lennard-Jones fluid, it is shown that the desired temperature distribution in each copy can be generated correctly in the long simulation. Not only physical properties can be estimated accurately, the convergence speed of them in the whole temperature range is enhanced as well compared to the CST method and the PCST method without exchange protocol. Especially for the low temperature copy, the introduction of exchange protocol ensures that it can sample various regions and minima in the configurational space, which will accelerate the convergence of physical properties.

PCST also shows its special effectiveness in protein folding simulation, in which the long residing time in low temperature range is required. In the application of folding trp-cage protein, we observed a significant enhancement of fraction of time in the native state compared to the CST method in each of the five
simulations. In addition, the adequate residing time in native states brings the observation of the existence of two substates in the native region. The molecular structures of two substates are in agreement with the previous computational\textsuperscript{30,31} and experimental\textsuperscript{31} studies. Moreover, three possible folding pathways are observed in the analyzed results of $C_\alpha$-$C_\alpha$ (helix) RMSD relationship, which indicates that various procedures might exist in the folding of trp-cage protein at certain conditions. The detection of two substates and multiple possible pathways demonstrates that PCST can sample the free energy landscape of protein folding thoroughly.

PCST method extends the ideas in CST method in using continuous tempering, while borrows the spirits in PT method in using multiple copies of simulation. The employment of various distributions in different copies overcomes the dilemma between barrier-crossing efficiency and importance sampling in particular temperature range, especially in the low temperature range. As the overlap between temperature distributions is sufficient, the exchange rate between copies can be kept at a relatively high value. At the same time, the exchange rate does not dramatically depend on the system size, which allows one to use much smaller number of copies compared to the conventional PT method. Therefore, it can serve as a good alternative for conventional PT in simulating large systems such as phase transition and dynamics of macromolecules in explicit solvent.
Conclusion

In this thesis, we first review the basic principles of statistical mechanics and traditional constant-temperature simulation methods in statistical mechanics. Generalized ensembles, which enhance the simulating efficiency over traditional methods, is reviewed in the second part of Chapter 2. Three major generalized ensembles: multi-canonical ensemble, simulated tempering and parallel tempering with their pros and cons, are also discussed. It is worth mentioning that the limitation of many computational methods emerges as the system size increase, which often makes the energy landscape even more rugged and complex.

In Chapter 3, we first summarize the continuous simulated tempering (CST) method. CST overcomes the issue of low acceptance ratio for large systems, which exists in both conventional simulated tempering and parallel tempering method. Next, an extension of CST into multi-copy simulation, which is parallel continuous simulated tempering (PCST), is proposed for even higher sampling efficiency in the important temperature ranges. Statistical methods such as MEII and adaptive averaging enable the simulated tempering methods to maximize the data use and accelerate the convergence of estimated physical properties. Therefore, PCST can serve as a good alternative of conventional parallel tempering in simulating large systems such as phase transition in molecular systems and dynamics of macromolecules in the explicit solvent.

In the future, we hope to extend our PCST method to many other applications and combine our sampling method with other approaches in the field of phase
transition simulation as well as the global optimization problems such as the protein structure prediction.
References