Fast and Accurate Enhanced Sampling

Application and Resource Efficiency for Replica Exchange Simulations with Temperature Intervals, Global Exchanges and Hybrid Solvent (TIGER2hs)

L. Schulig, M. Kulke, N. Geist, A. Link, M. Delcea,
Institute of Biochemistry & Institute of Pharmacy,
University of Greifswald

In Short

- Optimizing the performance of TIGER2hs – a replica exchange algorithm
- Determination of the optimal replica number for different peptides
- Definition of an empirical formular for the approximation of the optimal replica number

Introduction Determining the secondary structure of protein domains is often accompanied by high costs in experimental Nuclear Magnetic Resonance or X-ray studies. Also sometimes, proteins cannot be measured or crystallized, because they are for instance intrinsically disordered or membrane bound. In these cases, modeling approaches that predict the structure from the amino acid sequence are a cost-efficient alternative for small peptides. Recently, we developed a fast and accurate method (TIGER2hs)[1] for prediction of the secondary structure of proteins and applied it successfully for predicting the conformation of the C-terminal cross linking region of fibronectin, a protein domain containing a total of 100 amino acids[2].

This novel accelerated enhanced sampling method may replace the original replica exchange molecular dynamics (REMD)[3] approach in many fields and may enable investigations in systems were it was previously not feasible. While REMD has been successfully applied in thousands of studies since it was presented in 1999, its low computational efficiency prohibits a larger use of this tool by the simulation community. Parallel MD simulations are utilized at different temperatures to allow exploring a molecular systems entire phase space and this is of tremendous interest. However, in the genuine REMD the system size dictates the number of replicas and thereby the required computational resources. Hence, REMD is barely applicable with accurate explicit solvent conditions for most systems of biological relevance, nowadays. For TIGER2hs[1] in turn, the number of replicas scales linearly and is five to ten times lower for small proteins.

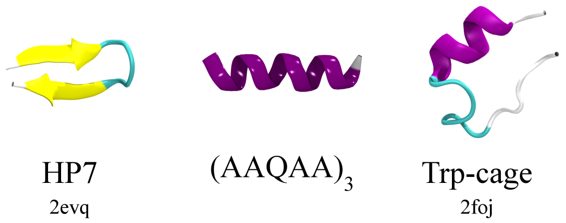

Figure 1: *Example peptides of different sizes used in this study (ordered by their size with 12, 15 and 20 amino acids, respectively).*

This is an advantage of TIGER2hs, because it allows the adjustment of the resources to the available computational systems. However, for very large folding projects this may lead to a problem. Here, available resources are not the limiting factor and it is more desirable to tune the sampling for best efficiency. While TIGER2hs has been shown to perfectly reproduce the thermodynamic ensemble of proteins, so far there is basically no available concept on how to select the number of walkers on the conformational landscape, but the performance crucially depends on the amount of replicas. Considering the extremes, a low number of two replicas may not yield much better sampling than with a single molecular dynamics simulation. On the one hand, although energy barriers may be overcome in the high temperature replica, the sampling is restricted to only a local ensemble and the exchange probabilities to the baseline replica – the replica with the lowest temperature – may be small. This results in a long time for convergence. With an increase in the number of replicas these effects have less and less impact on the convergence time and this time decreases. On the other hand, there may be a very large number of replicas. Increasing this number of replicas is more efficient than simply extending the simulation time, as long as it leads to the equivalent reduction in simulation time until convergence is reached. Naturally, the number of replicas cannot be increased to infinity and this means that for every folding simulation, there has to be an optimal replica cost that provides the best efficiency in terms of resource usage.

Objective In this project, we aim to investigate the best efficiency by determining the optimal number of replicas to use in a TIGER2hs[1] simulation as a function of the solute's degrees of freedom. To achieve this, replica exchange simulations with differ-

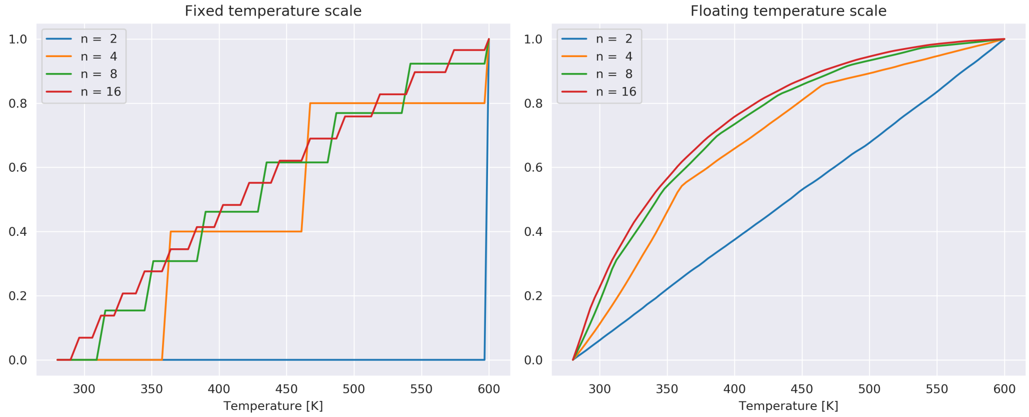

Figure 2: *Cumulative temperature distribution for a fixed temperature scale compared to the new floating maximum temperature approach.*

ent numbers of replicas are performed for fast-folding proteins of different sizes (Figure 1). This optimum is defined as the smallest use of resources where the correct average ensemble is obtained within the shorter convergence time in respect to the number of replicas. We assume that this optimal configuration is majorly dependent on the length of the amino acid sequence and an empirical function that allows the approximation of the optimal replica number for also larger systems may be defined. The gained insight into the scaling behavior and convergence will be required to open up the TIGER2hs tool for larger folding applications of protein modules with over 100 amino acids, where to date no accurate sampling was affordable.

Improved exchange algorithm A revision of the multiple replica exchange unveiled major problems. A state was written only when it got accepted as the new baseline replica and not necessarily as a result of the Metropolis criterion and therefore distorts the ratio in the ensemble. With large number of replicas, many simulations are at lower temperatures and the chance to find, keep and accept the native or an energy favorable structure increases. As obtained by our simulation results, these are overestimated with increasing replica number since the exchange is much more likely. To solve these issues, we implemented a single exchange algorithm with a random choice of replica.

Floating maximum temperature From what we know so far, one of the main reason for a minimum number of replicas is, that it is possible to heat up an already folded structure directly from minimum to maxmimum temperature which would be impossible in meaningful T-REMD simulations with more than two replicas. The time to fold a protein is much larger than an unfolding at high temperatures. To compensate this problem, we introduced a floating maximum temperature algorithm, where the maximum temperature is selected for every run by a random number on the initial temperature scale. Then, all replicas are logarithmic scaled based on the chosen maximum temperature.

WWW

https://biochemie.uni-greifswald.de/forschung/forschung-in-den-arbeitskreisen/ordner-aks-lehrstuehle/biophysikalische-chemie

More Information

[1] Geist, N.; Kulke, M.; Schulig, L.; Link, A.; Langel, W., *J. Phys. Chem. B* **2019**, *123*, 5996-6006.

[2] Kulke, M.; Uhrhan, M.; Geist, N.; Ohler, B.; Brüggemann, D.; Langel, W.; Köppen, S., *J. Chem. Inf. Model* **2019**, *59* (10), 4383-4392.

[3] Sugita, Y.; Okamoto, Y., *Chem. Phys. Lett.* **1999**, *314* (1-2), 141-151.