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

***M. Kulke, N. Geist, L. Schulig, M. Delcea**, Institut für Biochemie, Universität Greifswald*

In Short

- Optimizing the performance of TIGER2hs - a replica exchange algorithm
- Determination of the optimal replica number for five different peptides
- Definition of an empirical formular for the approximation of the optimal replica number based on the sequence length

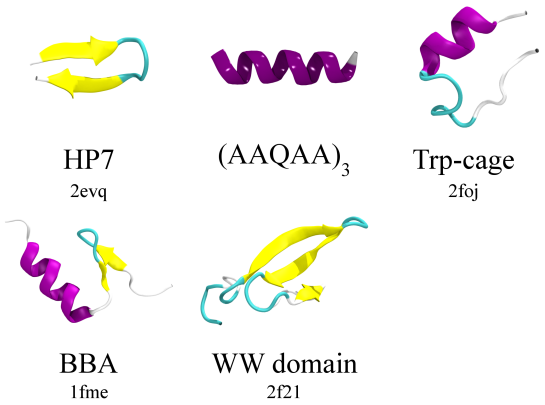

***Figure 1:** The five small peptides of different sizes used in this study. They are ordered by their size with 12, 15, 20, 28 and 39 amino acids, respectively.*

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 predicting 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[1]. 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[2] in turn, the number of replicas can be chosen freely.

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.

In this project, we aim to investigate the best efficiency by determining the optimal number of replicas to use in a TIGER2hs[2] simulation as a function of the amino acid sequence length. To achieve this, replica exchange simulations with 2, 4, 8, 16 and 32 replicas are performed for five fast-folding proteins of different sizes (Figure 1). The best performance is defined as the smallest product of replica num-

ber and convergence time for the simulation. 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.

The simulations shall proceed in five steps:

Step 1: Replica exchange simulations of HP7. Hp7 is a 12 amino acid small peptide that forms an antiparallel sheet (PDB code 2evq; Figure 1)[4]. Nevertheless, the structure is highly flexible and steadily changes between folded and unfolded having a fraction folded of ca. 8 %. The TIGER2hs algorithm previously folded the peptide with 8 replicas and great success resulting in the correct Boltzmann ensemble[1]. The convergence time for this peptide is around 400 ns and the correct determination requires at least 1 µs sampling time.

With a series of replica exchange simulations with different replica numbers, the most efficient computational resource use will be determined for this peptide. We expect the optimal replica count to be between 2 and 32 replicas.

Step 2: Replica exchange simulations of $(AAQAA)_3$. The protein $(AAQAA)_3$ forms a single helix and is often used in replica exchange studies. The maximum helix probability was experimentally found to be 45 %[5], while the folding studies estimated 30 %. Despite this discrepancy, the folding algorithms are consistent compared to each other[1][6]. The time of convergence is around 150 ns and for the correct estimation at least 300 ns sampling time is needed. This already suggests that convergence times differ significantly between peptides and especially for peptides containing sheet-motifs, the usage of the correct replica number is important to save computational resources. Again, with multiple replica exchange simulations varying the number of replicas, the best number will be derived. The expected optimal count is between 2 and 32 replicas.

Step 3: Replica exchange simulations of Trp-cage. This peptide is very similar to $(AAQAA)_3$ with the difference that a structurally defined coil sequence is attached[7]. The convergence times and optimal number of replicas is considered similar to $(AAQAA)_3$.

Step 4: Replica exchange simulations of BBA. BBA contains an helix as well as an antiparallel sheet and is therefore more challenging than the previous peptides[8]. The expected convergence time is in between the times for HP7 and Trp-cage resulting in roughly 600 ns sampling time needed, and the optimal replica count is similar to HP7.

Step 5: Replica exchange simulations of the WW domain. This peptide with 39 amino acids is the most difficult to fold but still in the capability of TIGER2hs[2]. It forms a large antiparallel sheet with a short random coil sequence at the N-terminus[9]. Due to the similarity to HP7, presumably 1 µs sampling time is needed for the accurate determination of the convergence time and the optimal replica count is presumably again between 2 and 32 replicas.

WWW

https://biochemie.uni-greifswald.de/forschung/forschung-in-den-arbeitskreisen/ordner-aks-lehrstuehle/biophysikalische-chemie-april-2019/

More Information

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

[2] Kulke, M.; Uhrhan, M.; Geist, N.; Ohler, B.; Brüggemann, D.; Langel, W.; Köppen, S., *Proteins Struct. Funct. Bioinforma.* **2019**, *submitted*.

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

[4] Andersen, N. H.; Olsen, K. A.; Fesinmeyer, R. M.; Tan, X.; Hudson, F. M.; Eidenschink, L. A.; Farazi, S. R., *J. Am. Chem. Soc.* **2006**, *128* (18), 6101-6110.

[5] Shalongo, W.; Dugad, L.; Stellwagen, E., *J. Am. Chem. Soc.* **1994**, *116* (18), 8288-8293.

[6] Kulke, M.; Geist, N.; Möller, D.; Langel, W., *J. Phys. Chem. B* **2018**, *122* (29), 7295-7307.

[7] Sheng, Y.; Saridakis, V.; Sarkari, F.; Duan, S.; Wu, T.; Arrowsmith, C. H.; Frappier, L., *Nat. Struct. Mol. Biol.* **2006**, *13* (3), 285-291.

[8] Plazzi, F.; Ribani, A.; Passamonti, M., *BMC Genomics* **2013**, *14* (1), 409.

[9] Clark, R. J.; Daly, N. L.; Craik, D. J., *Biochem. J.* **2006**, *394* (1), 85-93.