EXTENSIVE EXPLORATION OF CONFORMATIONAL SPACE IMPROVES ROSETTA RESULTS FOR SHORT PROTEIN DOMAINS

Yaohang Li

Department of Computer Science, North Carolina A&T State University
Greensboro, NC 27411, USA
Email: yaohang@ncat.edu

Andrew J. Bordner, Yuan Tian, Xiuping Tao, and Andrey A. Gorin*

Computer Science and Mathematics Division, Oak Ridge National Laboratory,
Oak Ridge, TN, 37831, USA
**Email: agor@ornl.gov*

With some simplifications, computational protein folding can be understood as an optimization problem of a potential energy function on a variable space consisting of all conformation for a given protein molecule. It is well known that realistic energy potentials are very "rough" functions, when expressed in the standard variables, and the folding trajectories can be easily trapped in multiple local minima. We have integrated our variation of Parallel Tempering optimization into the protein folding program Rosetta in order to improve its capability to overcome energy barriers and estimate how such improvement will influence the quality of the folded protein domains. Here we report that (1) Parallel Tempering Rosetta (PTR) is significantly better in the exploration of protein structures than previous implementations of the program; (2) systematic improvements are observed across a large benchmark set in the parameters that are normally followed to estimate robustness of the folding; (3) these improvements are most dramatic in the subset of the shortest domains, where high-quality structures have been obtained for >75% of all tested sequences. Further analysis of the results will improve our understanding of protein conformational space and lead to new improvements in the protein folding methodology, while the current PTR implementation should be very efficient for short (up to ~80 a.a.) protein domains and therefore may find practical application in system biology studies.

1. INTRODUCTION

The Rosetta platform[1-4] is one of the most successful approaches in predicting overall backbone fold for the protein domains that lack any detectable structural analogs in Protein Data Bank (PDB). It has been ranked number one at the last three CASP competitions (Critical Assessment of Structure Prediction) among *ab initio* methods[5]. Unlike threading methods that rely on a known structure template, ab initio programs attempt to predict structure by generating polymer chain configurations from the whole conformational space and use scoring functions to estimate how good these conformations are.

The Rosetta approach combines many innovative ideas to overcome the enormous complexity of the protein chain conformational space. Two of the most important features are: (a) fragment libraries and (b) knowledge-based energy potentials derived from the statistical analysis of known conformations. The fragment libraries contain custom-made lists of conformers for 3-mer and 9-mer segments centered on each residue of the target chain. This arrangement replaces more traditional polymer chain representations (e.g. by dihedral angles or Cartesian coordinates of the atoms) with a set of discrete variables – numbers of the conformers from the fragment library – with each of them determining the structure of the whole short segment of the chain. The segment libraries reduce the dimensionality of the conformational space by many orders of magnitude, however, for a chain of 200 residues it is still ~200 dimensions to explore. The conformations are evaluated based on their backbone atoms, as all side groups are replaced with "elastic spheres" and not modeled explicitly.

Rosetta operates by starting 1,000 (in latest implementations sometimes 10,000 or even more) independent folding trajectories from random extended conformations and evolving them with a Monte-Carlo procedure, while gradually reducing the temperature. For each trajectory, the structure with the lowest observed energy is retained as the result of the folding, and the corresponding 1,000 (or more) results are further analyzed by various methods to determine the

* Corresponding author.

native fold. We will not be discussing the computational problem of finding the native fold, as our study is concerned with the folding trajectories and the quality of the ensemble of the resulting backbone conformations. We will demonstrate that introducing parallel tempering dramatically improves sampling properties of the method and leads to better final structures, but the same results suggest that there are other problems in the procedure preventing more complete success.

2. METHOD

The Parallel Tempering algorithm[6-8] (also known as the multiple Markov chains or replica-exchange method) allows multiple Markov chains to evolve at several temperature levels, forming a ladder, while replica exchanges are attempted between Markov chains at neighboring temperature levels. We have introduced a few modifications to the PT algorithm without changing its fundamentals[9].

A composite system is constructed with one molecule per temperature level and the Rosetta-style transitions take place in each Markov chain. However, instead of the Simulated Annealing[15] scheme used in Rosetta, we use an adaptable Metropolis[14] scheme that maintains a desired acceptance rate. The replica exchange transition takes place according to the Metropolis-Hastings criterion. The desired acceptance rate is decreased gradually to accelerate convergence of the composite system[10]. Moreover, in protein modeling, each replica configuration consists of a lot of information and thus the exchange of configurations is very costly. Alternatively, we exchange the temperatures of two neighbor levels instead to achieve a significant computational performance improvement[11]. The topic of the conformational sampling in protein folding is explored in many excellent stidues[16-25], our investigation was limited to specific issues of the Rosetta folding platform.

We have followed Rosetta methodology and generated an ensemble of 1000 structures for each of 50 domains that were included in this study and each folding experiment. Several types of folding experiments were conducted: the usual Rosetta folding (further referred to as a Rosetta run) with 32,000 Monte-Carlo steps, PTR folding (in the figures referred to as an MPI run as the MPI library was used for multiprocessor implementation) with the same 32,000 steps during the main simulation stage, as well as the PTR runs with 320,000 steps (LMPI - Long MPI), and the PTR runs with $1.5 \cdot 10^6$ steps (referred to as a VLMPI or Very Long MPI). Rosetta was outperformed in MPI runs without additional CPU costs, because the final structure was collected from each thread in the PTR simulations. Due to certain CPU time restrictions only the LMPI protocol was done for all 50 tested domains, and these are the best results that we currently have. Table 1 and Fig. 1 are based on the LMPI protocol.

All modifications made to the original Rosetta package were limited to the sampling procedure. Rosetta records all parameters of the conformation with the lowest energy and (if the native structure is provided) the Minimal Root Mean Square Deviation (MRMSD) distance to the native structure over all structures observed during the simulation. This distance is often smaller than the RMSD distance between the final lowest energy structure and native model, but it is a good measure of how close to the native structure we were able to "pass" during the simulation.

3. RESULTS

3.1. Capability of traversing a "rough" energy landscape

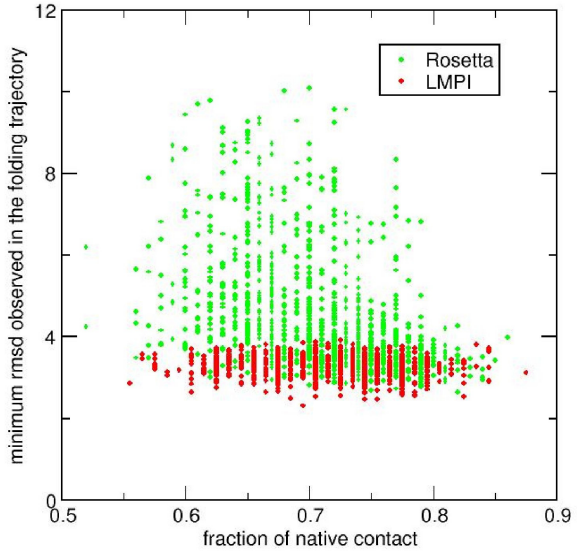

Fig. 1. Comparison of MRMSD to the fraction of native contacts in the final structure (Y-axis) for two ensembles of Rosetta and Parallel Tempering Rosetta simulations. All PTR trajectories pass within 4 Å RMSD of the native structure. Each point combines information from two different conformations, so there is no direct correlation between X and Y values.

All achieved improvements in the folding performance can be traced to the novel feature of Parallel Tempering Rosetta: the capability to traverse the rough energy landscape and get out from very deep local minima of the potential. In Fig. 1 two structure ensembles (each ~1000 structures) present results obtained for a Rosetta run (grey dots, wide area) and an LMPI run (darker dots, spread on much smaller area). The Y-axis represents the measure of the closest observed approximation to the native structure for a given trajectory — Minimal Root Mean Square Deviation (MRMSD) in Angstroms (Å). The X-axis displays the Fraction of Native residue Contacts (Cb-Cb under 8 Å) in the final structure for the corresponding trajectory. We know both of those quantities because we deal with a benchmark set, where the native structures are known.

There is a remarkable compression along the vertical axis. Only ~10% of all original trajectories have approached the native structure to the distance of 4 Å RMSD, but *all* 1000 trajectories in the PTR runs have passed below this limit. Actually, almost all of them have passed below 3 Å, with several trajectories reaching toward the 2 Å limit ("crystallographic" vicinity of native structure). It is important to note than any improvements in MRMSD is exponentially hard, as

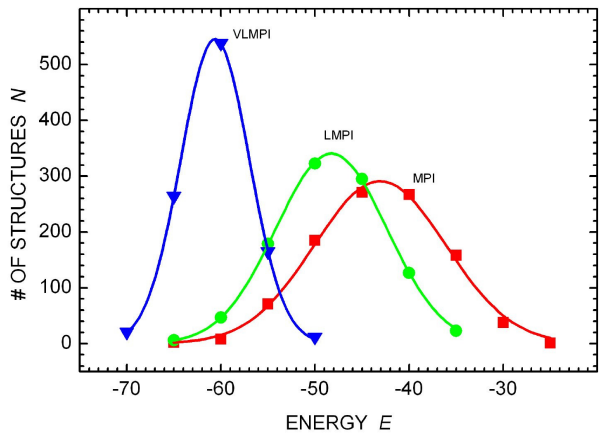

Fig. 2. Energy distribution of the final structures for 3 PTR runs: 32,000 step (MPI), 320,000 steps (LMPI) and $1.5*10^6$ steps (VLMPI). The effect of observing so many new conformations due to longer simulations has never been seen before for the Rosetta program.

the conformational volume shrinks very fast when one considers smaller and smaller RMSD "volumes". For example, in the Cartesian coordinates representation the conformational volume of the structures within a 2 Å RMSD vicinity of the native one is at least 8 times smaller than those in a 4 Å RMSD vicinity. Table 1 confirms that the results were typical for almost all analyzed domains, as in almost all cases we observed dramatic improvements in MRMSD.

Fig. 2 gives the most direct evidence that Parallel Tempering Rosetta reaches into new areas of the conformational space that could not be explored with standard simulated annealing Monte Carlo. The plot presents three energy distributions of the 1,000 final structures obtained in MPI, LMPI and VLMPI runs for the 1lev domain. The VLMPI run produces a much sharper distribution (twice as narrow), and it has little overlap with the MPI run. Here the lower energy is our "marker" that we indeed observe a novel conformation (Rosetta registers the lowest energy conformation seen).

As the distributions of the lowest energy visited by a particular trajectory show, it is clear that almost a half of the VLMPI runs have found conformations that were almost never visited by any of the MPI runs.

In the original Rosetta run, 32,000 steps used in the standard protocol were selected as the limit, after which there were no expectations of any improvements in the energy of the model. Here we observe an explosion in new conformations after extending the length of the run by 10 times (LMPI) and then by 5 times (VLMPI). In VLMPI case we even observe a semblance of convergence, as the width of the energy distribution starts to narrow. Interestingly, dramatic improvements in the final energy did not lead to equally dramatic improvements in the quality of the folded structures.

3.2. Results for the shortest domains

While improvements in the quality of the predictions have been seen across the whole benchmark, the simulations have reached a crucial "improvement threshold" for the shortest domains. The detailed results for the 16 shortest unique domains are presented in Table 1. In the original Rosetta run, the folding results are also systematically better for the shortest domains. With LMPI PTR simulations, several structures have been improved further, pushing the rate of good predictions to 75% of the total set in this size range (31 to 78 amino acids).

For 10 domains, the MRMSD parameter is under 2.5 Å (lines are shown in **bold** in Table 1). This means that at least one of the simulated trajectories passed within the *crystallographic quality vicinity* of the native structure (the corresponding numbers are underlined in

the table). Excellent final structures were found for all of them. Out of the remaining 7, three had MRMSD in the range of 3 to 4 Å with relatively good quality final structures.

Only for 4 structures (the whole lines underlined in Table 1) did our platform fail to find structures with the percent of native contacts much above 40% (MRMSD was in the range of 5 to 6 Å). Yet those structures have shown some MRMSD improvements with longer simulation times. Between the MPI and LMPI runs the MinRMSD parameter has improved by 0.5 - 1 Å for four sequences. Actually in this whole set MRMSD did not improve for only four structures, which already had excellent prediction quality by the original Rosetta program. Overall, a higher rate of success than ours has never been reported, to our knowledge, in the literature before. Further experiments conducted in our group confirm this result on a much larger set of unique sequences. Initial results on homologous sequences (the idea was to fold with Rosetta homologous domains as well) have indicated further improvements in two of the four "hard" sequences, pushing the overall success rate even higher.

3.3. Insights into the protein folding process

The conducted simulations and significantly improved ability to search conformational space led to important insights into the obstacles that are faced in computational protein folding. Fig. 3 plots the dependence between the length of the folded domains and the maximum fraction of native contacts (100 means an ideal native structure) obtained in one of the accepted models for this domain. To iron out the structural differences, we used "sliding window" averages for both coordinates (each point represents averages over 10 structures close in length). The results for 50 folded domains produce 41 "sliding windows", and the corresponding 41 points are presented in Fig. 3.

The dependence is sharp and non-linear — for a domain length of around 110 the fraction of native contacts is projected to be only around 30%. At this level there are probably some correct elements of secondary structure, but likely no correct tertiary contacts. The good news is that the results are close to excellent for the domains <75 residues. Another encouraging point is that the problems, which rapidly escalate with increasing the length of the polymer chains, are probably tractable by applying more computer power. Indeed, we have observed the largest amplitude of improvements measured by the fraction of native contacts in the final structure in the longest considered domains (L>90), when we extended simulation from MPI to LMPI protocol.

Table 1. The results for 16 domains in range of 31 to 78 amino acids. The domains are shown by PDB is and chain identifier.

Structure ID	Best final RMSD (A)	Best MRMSD observed (A)	Best final FNC (%)
1tgz_B	**3.3**	**1.81**	**81**
1r0o_B	8.7	6.11	40
2bf8_B	**3.0**	**1.80**	**74**
1xt9_B	**3.6**	**2.07**	**69**
1r0o_A	5.8	5.25	41
1sv0_A	**3.0**	**2.01**	**74**
1le8_B	**6.1**	**2.31**	**87**
1dj7_B	7.9	5.95	40
1oey_A	5.6	5.29	43
1cf7_A	**2.8**	**1.87**	**78**
1bun_B	4.4	3.61	49
1le8_A	**1.4**	**0.82**	**96**
4sgb_I	4.5	4.32	41
1nql_B	4.3	2.60	54
1j2j_B	**1.4**	**0.61**	**99**
1mzw_B	**2.2**	**1.06**	**85**

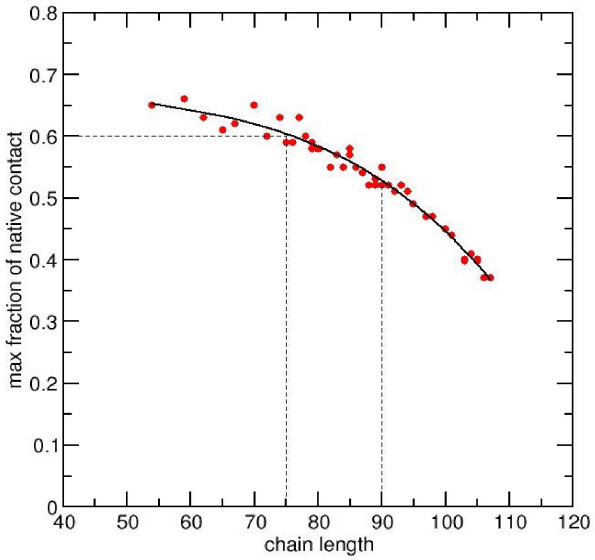

Fig. 3. The dependence between the length of the chain and the quality of the final structures. Below 75 amino acids the quality is very good, but it drops down sharply for longer domains.

The curve in Fig. 3 clearly spells trouble for Rosetta simulation of the domains longer than 105 residues. The average fraction of native contacts was only around 35% for domains in this range, and therefore correct folds can be expected only as a result of extraordinary luck.

4. DISCUSSION

In this study MRMSD measuremets have been used to access improvements in the capability to explore conformational space. Indeed, as the starting conformations are random elongated chains, during normal Rosetta simulations many of them never will fold successfully, and many simulation trajectories will never even pass in a close proximity of the native conformation. The fraction of the trajectories, which have conformation on a certain distance from the native conformation, is then an indirect indication of relative "freedom to travel" shown by the algorithm.

There are several important properties of the MRMSD that should be mentioned here. First, as we already mentioned above, the reduction of the "conformational volume" (defined in a reasonable metric it is simply a real volume in the space of conformational variables) is a power function of the reduction of the RMSD value. One can speculate that a reduction of the RMSD 2 times translates into 8 (or 16?) times reduction in the available conformational volume. Second, the MRMSD depends on the size of the protein chain in a complex way. For longer chains much smaller fraction of all configurations will satisfy the RMSD constraint of 2 Å than for shorter ones. Finally, even a very good MRMSD value does not guarantee that the folding will be successful. The structural trajectory will include a conformation with a 2 Å MRMSD value, but this conformation may have a high value of potential energy (due to some highly unfarobale interactions present in the overall correct model). As a result, the candidate conformation will not be saved, and in the following simulation the final conformation will be very different.

On other hand if a particular folding trajectory does not show a good value of the MRMSD than the simulation is bound to be unseccesful. Due to the definition an MRMSD value of, for example, 8 Å means that the best RMSD possible for the final structure will be greater or at best equal to 8 Å. This simple point explains our efforts to achieve a good MRMSD value for *all* folding trajectories. The trajectories with bad MRMSDs are essentially waste of the CPU time.

To access the quality of the resulting structures we have used (in addition to a standard RMSD) another measure: Fraction of Native residue Contacts (FNC). Two residues were considred to be in contact if the distance between their Cb atoms (Ca for the glycin) was smaller than 8 Å. The "automatic" contacts (with neigbours -2, -1, +1, +2) were excluded. Many possible definitions of contacting residues are possible, for example, one can define differential contact cutoffs to take into account residue size differences.

By our experience almost all reasonable FNC definitions work well, and there is no clear advantages to prefer one definition to another. For some types of the analysis it seems to be useful to distingvish between short-range (local) and long-range contacts. The long-range contacts provide a more sensitive measure of the folding success, but then there is an additional uncertenty due to the noise effect, which is stronger on smaller sets of contacts. The FNC may provide a superior measure for the quality of the folded structures, but the questions about relative contributions of local and long-range contacts deserve a separate investigation. One possible way forward would be to use weights on all contacts derived, for example, from the separation between contacting residues in the primary sequence.

In the future we plan to conduct more comprehensive analysis of the folding trajectories.

Currently for each trajectory only two (most important) trajectory points are recorded: the conformation with the lowest energy (for this one we have the full set of data) and the lowest RMSD distance to the native fold (here we are limited to the distance value). Nevertheless several interesting and important conclusions both practical and theoretical can be drawn from the current work.

First, the Parallel Tempering dramatically improves sampling capabilities of the program. All local minima can be comprehensively explored. In the longest simulations we have observed an emerging Monte-Carlo convergence of the trajectories. Here we should note that these results obtained on relatively "soft" potentials. The real energy potentials (such as electrostatic and Van der Waals interactions) usually lead to rougher potential energy functions that the knowledge-based derived potentials. Yet there is no reason to believe that the Parallel Tempering algorithm cannot be adapted to such potentials with more temperature energy levels, etc.

Indeed, the role of the potential energy function constitutes a second lesson of our study. In a number of situations we observed that the current potential functions lead to a large "valley", where the native structure is located, but this valley does not have deep potential energy minimum located at the native conformation. While almost all folding trajectories cross the right "valley", only very few of them end up near the native conformation. There is no energy gradient leading through the remaining 2 Å of RMSD — and this process happens almost randomly, increasingly so for longer domains. Our approach will be useful for a more detailed exploration the conformational space and properties of the potentials. For example, we can produce structures with very low values of potential energy, which are really far from the native model, and in such way reveal shortcomings of the existing potentials.

The final (and helpful for applications) conclusion from our study is a sharp dependence between the probability to have a successful folding result and the length of the targeted domain (presented in Fig. 3). For short domains (75-90 residues long) the PTR implementation provides a significant improvement over the standard Rosetta, with high chances to have a structure with 80% of native contacts in the final ensemble. This improvement is something like making of the "last mile" for the folding, because the original Rosetta is also pretty good for such short domains.

On a separate topic we note that the identification of the best native candidates (something we do not explore in this paper) will be facilitated by the PTR property mentioned above. Almost every trajectory will be drawn into the "valley" around the native structure, so if the near native state tends to be occupied, many more near native decoys will be produced with the PTR than with usual Monte-Carlo simulated annealing Rosetta.References

Acknowledgements

The study was supported by the LDRD Program of the Oak Ridge National Laboratory managed by UT-Battelle, LLC, under Contract DE-AC05-00OR22725

References

1. Simons KT, Bonneau R, Ruczinski I, and Baker D. Ab initio Protein Structure Prediction of CASP III Targets Using ROSETTA, *Proteins: Structure, Function and Genetics*, 1999; **37**: 171-176.
2. Baker D. A surprising simplicity to protein folding, *Nature*, 2000; **405**: 39-42.
3. Bradley P, Chivian D, Meiler J, Misura KM, Rohl C, Schief W, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation, *Proteins*, 2003; **53**: 457-468.
4. Rohl CA, Strauss CE, Misura KM, Baker D. Protein Structure Prediction using Rosetta, *Methods in Enzymology*, 2004; **383**: 66-93.
5. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction, *Curr. Opin. Struct. Biol.*, 2005; **15**: 285–289.
6. Geyer CJ, Thompson EA. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference, *Journal of the American Statistical Association*, 1995; **90**: 909-920.
7. Hansmann U. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules, *Chem. Phys. Letter*, 1997; **281**: 140-150.
8. Li Y, Strauss CE, Gorin A. Parallel Tempering in Rosetta Practice, *Proceedings of International Conference on*

Bioinformatics and its Applications, Fort Lauderdale, Florida, 2004.

9. Li Y, Protopopescu VA, Gorin A. Accelerated Simulated Tempering, *Physics Letters A*, 2004; **328**: 274-283.
10. Li Y, Strauss CE, Gorin A. Hybrid Parallel Tempering and Simulated Annealing Method – an Efficient Sampling Method in ab initio Protein Folding, *International Journal of Computational Science*, 2008; in print.
11. Li Y, Mascagni M, Gorin A. Decentralized Replica Exchange Parallel Tempering: An Efficient Implementation of Parallel Tempering using MPI and SPRNG, Proceedings of International Conference on Computational Science and Its Applications (ICCSA), Kuala Lumpur, 2007.
12. Schug A, Herges T, Verma A, Wenzel W. Investigation of the parallel tempering method for protein folding, *J. Phys: Condens. Matter*, 2005; **17**: 1641-1650.
13. Schug A, Wenzel W. Predictive in-silico all atom folding of a four helix protein with a free energy model, *J. Am. Chem. Soc.*, 2004; **126**: 16737.
14. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller TH, Teller E. Equation of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 1953; **21**: 1087-1092.
15. Kirkpatrick S, Gelatt DDJ, Vecchi MP. Optimization by Simulated Annealing, *Science*, 1983; **220**: 671-680. Hansmann U. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules, *Chem. Phys. Letter*, 1997; **281**: 140-150.
16. Hansmann U. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules, Chem. Phys. Letter, 1997; 281: 140-150.
17. Lin C, Hu C. Hansmann UHE. Parallel Tempering Simulations of HP-36, *Proteins, Structure, Function, and Genetics*, 2003; **52**: 436-445.
18. Rabor AA, Scheraga. Improved Genetic Algorithm for the Protein Folding Problem by Use of a Cartesian Combination Operator, *Protein Science*, 1996; **5**(9): 1800-1815.
19. Pedersen JT, Moult J. Protein Folding Simulations with Genetic Algoithms and a Detailed Molecular Description, *J Mol Biol.*, 1997; **268**(2):240-259.
20. Damsbo M, Kinnear BS, Hartings MR, Ruhoff PT, Jarrold MF, Ratner MA, Application of evolutionary algorithm methods to polypeptide folding: Comparison with experimental results for unsolvated Ac-(Ala-Gly-Gly)5-LysH+, *Proceedings of the National Academy of Sciences*, 2004; **101**(19): 7215-7222.
21. Schulze-Kremer S. Application of Evolutionary Computation to Protein Folding, *Advances in evolutionary computing: Theory and Applications*, 2003; 915-940.
22. Kim JG, Fukunishi Y, Nakamura H. Average Energy Guided Simulated Tempering Implemented into Molecular Dynamics Algorithm for Protein Folding Simulation, *Chemical Physics Letters*, 2004; **392**: 34-39.
23. Okamoto Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations, Journal of Molecular Graphics and Modelling, 2004; **22**: 425 - 439.
24. Mitsutake A, Sugita Y, Okamoto Y. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides, *Journal of Chemical Physics*, 2003; **118**: 6664 - 6675.
25. Sugita Y, Okamoto Y, Replica-exchange molecular dynamics method for protein folding, Chemical Physics Letters, 1999; 314: 141-151.