PARALLEL TEMPERING IN ROSETTA PRACTICE

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

CHARLIE E. M. STRAUSS
Computer Science and Mathematics Division
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
email: cems@lanl.gov

ANDREY GORIN
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
email: agor@ornl.gov

Parallel Tempering (PT) is an effective algorithm to overcome the slow convergence in low-temperature protein simulation by initiating multiple systems to run at multiple temperature levels and randomly switch with neighbor temperature levels. We implemented the PT scheme in the Rosetta to explore the rough energy landscape in protein folding and to improve the success rate of Rosetta in topologically complex structures. Compared to the original Simulated Annealing (SA) scheme in Rosetta, our preliminary computational results show that the PT scheme in Rosetta program exhibits a wider range sampling in the potential energy surface in protein folding.

1. Introduction

One of the most important open problems in molecular biology is the prediction of the spatial conformation of a protein from its primary structure, i.e., from the linear sequence of its amino acids. Structural genomic enterprise crucially depends on the predictive tools to extend knowledge from the studied species to those where the corresponding proteins cannot be explored with the direct methods. The Rosetta program [1, 2] currently is the leading tool in predicting overall backbone fold for the protein domains that lacks any detectable structural analogs in PDB.

The Rosetta algorithm employs a Simulated Annealing (SA) method [3] to produce a Markov process for exploring the huge conformational space of amino acid sequences with the goal finding the conformation with the global energy minimum. The SA method is based on the fact that it is easier for a system to escape from a local minimum energy at a significantly higher temperature. Thus, within the SA scheme, the temperature is initially raised to a high value. The system evolves according to the Metropolis-Hastings criterion [7], while the temperature is very slowly reduced during the simulation procedure.

The energy landscape of protein folding usually has an intricate surface with high degree of freedom. Our research [8] shows that the SA method will be easily trapped in deep local minima in a simulated rough “funnel” energy function. This also indicates that, in the energy landscape of protein folding that usually roughly resembles a funnel with hierarchically disposed local minima [9], the sampling process may also be unable to escape from the deep local minima.

Compared to the SA method, the Parallel Tempering (PT) method [6, 10, 11] has multiple systems, one at each temperature level, and enables the system at the low temperature level to escape from local minima and to locate multiple minima by allowing the system to switch with the system configuration at higher temperature ladder according to the Metropolis-Hastings rule. Recent researches [11, 12, 13] indicate that, with respect to obtaining low temperature system configurations, PT is superior to simple Monte Carlo and to SA.

In this paper, we will adopt the Parallel Tempering scheme in the Rosetta program to explore the complex protein folding landscape. The remainder of this paper is organized as follows. We illustrate the overview of Rosetta program and the Rosetta PT algorithm in Sections 2 and 3, respectively. In Section 4, we discuss the preliminary results of applying the PT algorithm to Rosetta program. Finally, Section 5 summarizes our conclusions and future research directions.

2. Rosetta Algorithm

The Rosetta program currently is the most successful tool in predicting overall backbone fold for the protein domains that lacks any detectable structural analogs in PDB. Rosetta combines many ideas for the acceleration of the structure search discussed above: collective variables, conformer libraries and empirical energy functions. The idea of collective variable approach is given an interesting turn in Rosetta. Each protein chain residue has two associated libraries of the conformers for 3-mer and 9-mer segments centered at this residue, therefore allowing to change cooperatively structure of the whole segment by changing a single discreet parameter – number of the conformer we want to test at this moment. The conformers libraries for each residue are custom made from all available PDB structures before Rosetta starts constructing model for a given protein chain based on two criteria: sequence similarity and predicted secondary structure. For example, if the residue is predicted to have probabilities 50%, 30% and 20% for being in α-helix, β-strand and loop region, correspondingly, local segment library will have 50% of α-helix fragments, 30% of β-strand and 20% of other extracted from PDB with sequences, which most closely match the given sequence surrounding this residue.

Rosetta energy (scoring) function is a complex combination of the functions that mostly utilize PDB-derived measures of probability for a given configuration to be native one. One of the most important components is the table of probabilities to observe specific types of residues into certain distances from each other. In such approach the scoring function is smoothly integrating many types of physical interactions contributing to preferred arrangements of residues and implicitly taking into account factors that could be of evolutionary origin. This scoring function is extremely efficient computationally and very “soft” as it significantly smoothes out underlying rough energy landscape.

In the process of simulation Rosetta switches from softer energy functions to more “stiff” ones simultaneously changing the nature of the conformational steps it is using. At the beginning, large segments are changing the conformation dramatically swinging in space remote segments of the chain, on later stages it is using smaller changes preserving significant part of already found contacts and engaging small incremental adjustments in the individual dihedral angles. As all other simulation program Rosetta gives not a single answer but a set of answers, and one of the largest remaining problems is to select a correct one out of top models as among the score value alone is not reliable indicator when we are discussing top bin of models. For protein domains of 100-150 amino acids Rosetta most often contains a correct model among five top score, if we define as correct model reproducing fold of the segment (2-3 A difference in RMSD of the backbone), and does so equally well for domains with novel never before observed folds. This is a remarkable achievement, which would be totally unthinkable even 5 years ago, but there is still a lot of space for improvement.

Currently Rosetta has a lower success rate on topologically complex structures [4, 5]. The frequency of occurrence of high contact order or many stranded beta sheets is lower in Rosetta than in the natural protein population: the topological complexity of a protein dramatically decreases our capability of sampling near native conformations.

Protein conformational space is so large that long term molecular dynamic modeling quickly becomes untenable at even modest sized proteins. A known route to acceleration of the system dynamic evolution is to allow local minimums on a potential surface to freely exchange, thus avoiding kinetic bottlenecks. It is feasible that Rosetta performance could be significantly improved by borrowing advanced techniques from Monte-Carlo sampling theory developed in applied math theory several last years.

At present Rosetta over samples (i.e. wastes effort) because it tends to predict the same structures multiple times in separate runs of the program. Such behavior strongly indicates local minima trapping problem. It should be possible to borrow from multiple temperature ensemble Monte Carlo techniques to optimally sample potential energy surface.

3. Parallel Tempering in Rosetta Program

We developed the Parallel Tempering scheme in Rosetta program. The Rosetta program employs various types of moves, including smooth moves and chuck moves by evaluating different scoring functions (the environment score and the pair score). PT is adopted in all moves. In Rosetta's PT implementation, the Markov chains can be realized with two sets of moves:

1. Local Monte Carlo moves at each temperature level. The transition probability only depends on the change of in scoring function $E(C_i)$, where C_i is the configuration of the system at temperature level *i*. The Metropolis-Hastings [7] ratio at temperature level *i* is computed by
   $$w_{Local}(C_i^{old} \to C_i^{new}) = e^{-\beta_i \Delta_i E} = e^{-\beta_i (E(C_i^{new}) - E(C_i^{old}))},$$
   where $\beta_i = 1/k_B T_i$ with k_B the Boltzmann constant and T_i the temperature at level *i*. The new state is accepted with probability $\min(1, w_{Local}(C_i^{old} \to C_i^{new}))$.
2. Exchange of systems between two neighbor temperature levels, *i* and *i+1.*
   $$C_i^{new} = C_{i+1}^{old}$$
   $$C_{i+1}^{new} = C_i^{old}.$$

The exchange is accepted according to the Metropolis-Hastings criterion with probability $\min(1, e^{-\beta_i E(C_j) - \beta_j E(C_i) - \beta_j E(C_j) + \beta_i E(C_i)})$ to maintain the detailed balance condition of every Markov process. In the Rosetta program, the system consists of a lot of information and thus the exchange of systems is very costly. Alternatively, we exchange the temperatures of two neighbor levels instead.

In the PT scheme of Rosetta, the moves at high temperature levels intend to have a wide range exploration of the system energy landscape with a higher acceptance rate while those at low temperature levels explores the local details. Altogether, the PT scheme is expected to provide an efficient approach to explore the rough protein folding energy landscape.

4. Preliminary Results

Protein		Original Rosetta		Rosetta PT	
Name	structure	rms	rms-min	rms	rms-min
1a32_	alpha	6.75	2.17	6.87	1.41
1lis_	alpha	15.62	12.30	15.48	10.46
1lz1_	alpha-beta	15.42	12.79	15.43	11.51
2ptl_	alpha-beta	9.63	6.62	9.01	5.86
1d3z_	alpha	7.63	4.94	7.70	4.58
1gvp_	beta	12.67	10.42	12.90	9.53
1cg5B	alpha	14.67	12.02	14.29	10.37
1danT	beta	11.70	9.50	11.57	8.82
1elwA	alpha	10.71	3.02	10.63	3.30
1ig5A	alpha	6.48	4.25	6.76	4.00
1louA	alpha-beta	12.34	9.72	12.30	9.30
1opd_	alpha-beta	11.94	10.17	11.89	8.26
1pcfA	beta	8.17	4.95	8.34	4.93
1tig_	alpha-beta	8.65	6.72	8.05	5.55
1tuc_	beta	9.43	7.12	8.97	5.89
2acy_	alpha-beta	12.11	10.05	11.90	9.23
5croA	alpha-beta	7.80	4.74	7.92	4.73

Table 1: *RMS and RMS-MIN Comparisons of Rosetta with SA Scheme (Original Rosetta) and Rosetta with PT Scheme in Various Protein Target Structures

[*] The computations are carried out on a IBM Linux cluster with 16 nodes, 1G Memory each node.

Table 1 shows the RMS and RMS-MIN comparisons of the original Rosetta program with SA scheme and the Rosetta with PT scheme in various protein target structures. The PT scheme employs 4 temperature levels, ranging from 1.0 to 4.0 and Table 1 compares the average RMS and RMS-MIN values of the first 1000 conformations found in Rosetta and Rosetta PT search. The RMS column reflects the RMS value of the accepted conformation after an ab initio Rosetta search, which indicates the closeness of this conformation to the native structure. The RMS-MIN indicates the lowest RMS value of a protein conformation that is found during the Rosetta sampling procedure, however, the search moved away from this structure due to some reason, e.g., the structure had some pathology like a clash even though from the RMS point of view, it may have been a good one.

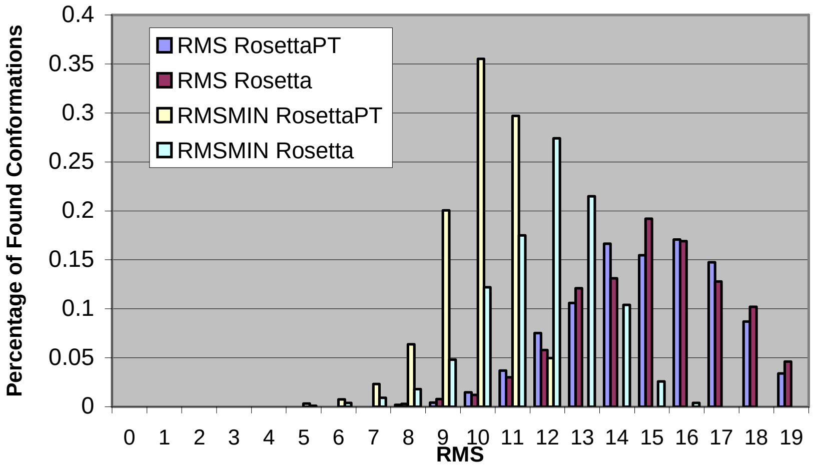

Figure 2: RMS and RMS-MIN Comparison of 1000 Conformations Found in RosettaPT and Original Rosetta in Protein Target Structure 1lis

The computational results shown in Table 1 indicate that Rosetta PT presents smaller RMS-MIN values in most experimental protein target structures than those of original Rosetta, most of which is 0.5 to 2.0 Å shift to 0 (native structure). Figure 2 shows the histogram of the RMS and RMS-MIN values of 1000 conformations generated by Rosetta and Rosetta PT in structure target 1lis. We can find Rosetta PT has a significant shift to the native structure in RMS-MIN from the original Rosetta using SA algorithm. This phenomenon implies that the Rosetta PT search procedure is “closer” to the native structure than the SA search used in original Rosetta. In other words, the PT scheme has a wider range of search in the protein folding energy landscape than the SA scheme in original Rosetta.

Nevertheless, the RMS values in Table 1 do not show a significant improvement in Rosetta PT. More clearly, Figure 2 and Figure 3, which shows the RMS and contact order of 1000 structure produced by Rosetta and Rosetta PT in 1lis, respectively, do not exhibit significant improvement in Rosetta PT scheme as well. Therefore, even though the PT scheme is "closer" to the native structure in the search procedure, due to the acceptance criteria in Rosetta, the "close" native structure is not actually accepted and the system then moves away.

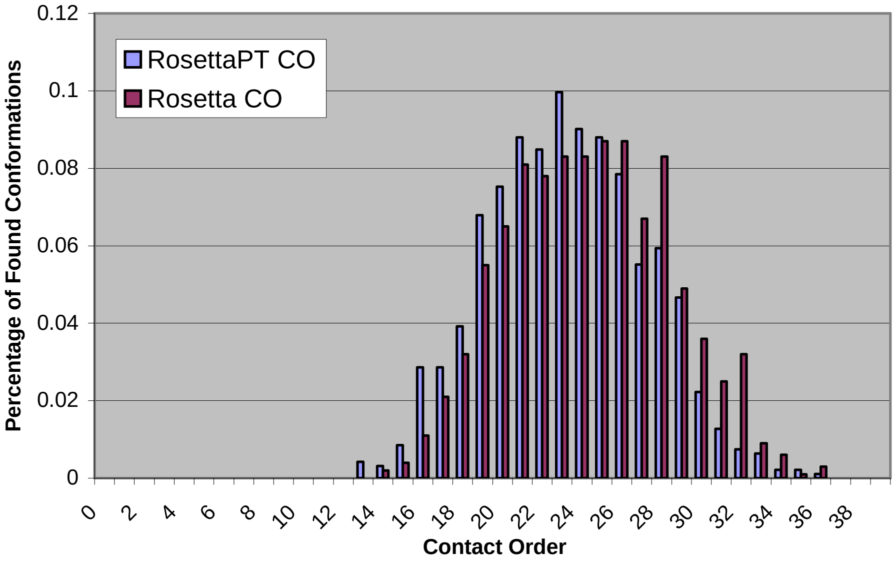

Figure 3: Contact Order Comparison of 1000 Conformations Found in RosettaPT and Original Rosetta in Protein Target Structure 1lis (The contact order of the native 1lis structure is 30.80)

To further analyze the PT scheme in Rosetta program, we take a closer look at the structures accepted in Rosetta PT at different temperature levels. Since the search algorithm in Rosetta actually explores the scoring function space, we display the scoring function and RMS values of the accepted structures in different temperature levels in Rosetta PT and those of the original Rosetta in Figure 4 and 5, respectively.

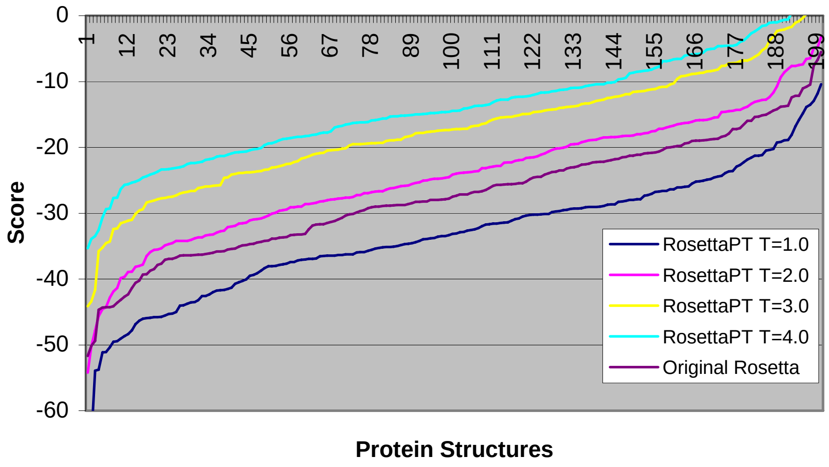

Figure 4: Sorted score values of structures at various temperature levels in Rosetta PT and those in original Rosetta at protein target 1lis. T is the temperature at each level.

Figure 4 shows the score values of the first 200 structures of protein target 1lis accepted at various temperature levels (T ranges from 1.0 to 4.0) in Rosetta PT and the score values of the first 200 structures accepted in original Rosetta. The score values are sorted in an ascending order. We can find that the conformations accepted at high temperature level exhibit averagely higher score than those at low temperature level and scoring function curve of original Rosetta locates between those of the high temperature level and the low temperature level in Rosetta PT. Figure 4 indicates that the Markov process in Rosetta PT at lower temperature level explores deeper minima in the scoring function in protein folding than that at higher temperature level.

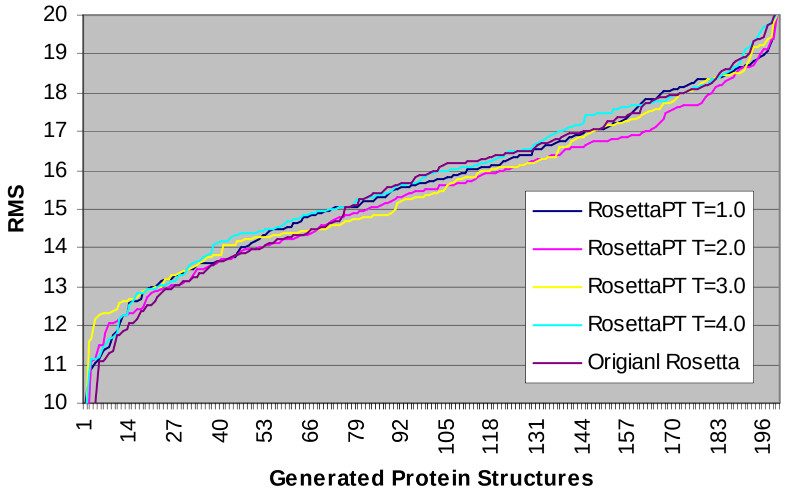

Figure 5: sorted RMS of structures at various temperature levels in Rosetta PT and those in original Rosetta

Figure 5 shows that the sorted RMS of the first 200 structures at various temperature levels in Rosetta PT and those in Rosetta. We can find that there is no significant difference in the RMS curves at different temperature levels in Rosetta PT and that of the original Rosetta. Both Figure 4 and 5 indicate that even though the structures accepted at lower temperature level in Rosetta PT yields smaller score values, the RMS of these structures are indistinguishable with those structures accepted at higher temperature. This explains why PT scheme in Rosetta is able to explore the wider range of the scoring function space and find the deeper energy minima, but the found structures in Rosetta PT does not have a significant RMS improvement. The reason is, low values in the scoring function do not indicate low RMS in the protein structure.

5. Conclusions and Future Research

The Parallel Tempering exhibits a fast convergence rate in complex molecular simulation. In this paper, we discuss applying the PT algorithm to Rosetta program with hope to have a more effective exploration in complex protein folding energy landscape. Our preliminary results show that the PT scheme in Rosetta program exhibits a wider range sampling in the scoring function surface. However, our results also indicate that even though the PT scheme has a wider range of exploration in the protein folding energy landscape, it does not show a significant improvement in RMS and contact orders of the generated protein structures in Rosetta PT scheme.

These preliminary computational results also raise several interesting research questions for us to explore:

1. Can we apply some other more effective sampling approaches, such as Accelerated Simulated Tempering (AST) [8], Accelerated Parallel Tempering (APT) [14], or dynamic weighting [15], to Rosetta program to achieve an even wider exploration of the protein energy landscape?
2. Can we adjust the acceptance criteria in Rosetta so that the structures that are "close" to native structures will not be ignored?
3. Can we also develop new scoring functions or refine the current scoring functions in Rosetta to more precisely reflect the RMS or contact order of a protein structure? Scoring functions based on the contact order may be a good choice.

6. Acknowledgments

We acknowledge partial support from the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under the project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling"; the LDRD Program of the Oak Ridge National Laboratory (ORNL); and the Division of Materials Science and Engineering, U. S. Department of Energy. ORNL is managed by UT-Battelle, LLC, under Contract DE-AC05-00OR22725.

7. References

1. K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, "Ab initio Protein Structure Prediction of CASP III Targets Using ROSETTA", Proteins: Structure, Function and Genetics, **37**(3): 171-176, 1999.
2. D. Baker, "A surprising simplicity to protein folding," Nature **405**: 39-42, 2000.
3. S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, "Optimization by Simulated Annealing," Science, **220**: 671-680, 1983.
4. K. W. Plaxco, K. T. Simons, D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins," J Mol Biol, **277**: 985-94, 1998.
5. R. Bonneau, I. Ruczinski, J. Tsai, D. Baker, "Contact order and ab initio protein structure prediction", Protein Sci, **11**(19): 37-44, 2002.
6. M. Falcioni, M. W. Deem, "A Biased Monte Carlo Scheme for Zeolite Structure Solution," *J. Chem. Phys*, **110**: 1754-1766, 1999.

7. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of State Calculations by Fast Computing Machines”, Journal of Chemical Physics, **21**: 1087-1092, 1953.
8. Y. Li, V. A. Protopopescu, A. Gorin, “Accelerated Simulated Tempering,” Physics Letters A, **328**(4): 274-283, 2004.
9. J. N. Onuchic, Z. Luthey-Schulten, and P. Wolynes, “Theory of Protein Folding: The Energy Landscape Perspective,” Annual Reviews in Physical Chemistry, **48**:545-600, 1997.
10. M. Falcioni, M. W. Deem, “A Biased Monte Carlo Scheme for Zeolite Structure Solution,” *J. Chem. Phys*, **110**: 1754-1766, 1999.
11. U. Hansmann, “Parallel Tempering Algorithm for Conformational Studies of Biological Molecules,” Chem. Phys. Letter, **281**: 140-150, 1997.
12. J. J. Moreno, H. G. Katzgraber, A. K. Hartmann, “Finding Low-Temperature States with Parallel Tempering, Simulated Annealing and Simple Monte Carlo,” Int. J. of Mod. Phys. C, **14**(3): 285-299, 2003.
13. C. Y. Lin, C. K. Hu, U. H. E. Hansmann, “Parallel Tempering Simulations of HP-36,” Proteins: Structure, Function, and Genetics, **52**(3): 436-445, 2003.
14. Y. Li, V. A. Protopopescu, A. Gorin, “Accelerated Simulated Tempering,” in preparation.
15. J. S. Liu, F. Liang, and W. H. Wong, “A Theory for Dynamic Weighting in Monte Carlo Computation,” Journal of the American Statistical Association, **96**: 561-573, 2001.