BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 14 2007, pages 1840–1842
doi:10.1093/bioinformatics/btm259

Structural bioinformatics

T-Pile—a package for thermodynamic calculations for biomolecules

Dominik Gront* and Andrzej Kolinski

Warsaw University, Faculty of Chemistry, Pasteura 1 02-093 Warsaw, Poland

Received on April 13, 2007; revised on April 13, 2007; accepted on May 18, 2007

Advance Access publication May 17, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Molecular dynamics and Monte Carlo, usually conducted in canonical ensemble, deliver a plethora of biomolecular conformations. Proper analysis of the simulation data is a crucial part of biophysical and bioinformatics studies. Sequence alignment problem can be also formulated in terms of Boltzmann distribution. Therefore tools for efficient analysis of canonical ensemble data become extremely valuable. T-Pile package, presented here provides a user-friendly implementation of most important algorithms such as multihistogram analysis and reweighting technique. The package can be used in studies of virtually any system governed by Boltzmann distribution.

Availability: T-Pile can be downloaded from: http://biocomp.chem.uw.edu.pl/services/tpile. These pages provide a comprehensive tutorial and documentation with illustrative examples of applications.

Contact: dgront@chem.uw.edu.pl

Supplementary information: Supplementary data are available at *Bioinformatics online*.

1 INTRODUCTION

Proper thermodynamic analysis of simulation results is crucial for understanding properties of biomolecules. Usually simulations are done in canonical or multicanonical ensembles. Canonical ensemble is more intuitive and easier to simulate. The drawback is that the partition function cannot be computed directly from a trajectory of states observed at a single temperature. Simulations at various temperatures can be combined by means of multihistogram algorithm (Ferrenberg and Swendsen, 1989; Gront *et al.*, 2001) [or WHAM Weighted Histogram Analysis Method (Kumar *et al.*, 1992)] introduced a few years ago. Despite the popularity of the method, there is no a publicly available program that implements it. The purpose of this work is to provide a flexible implementation that can be used in many applications.

2 T-PILE PACKAGE

The package has been implemented in Java language what makes it almost platform independent. Special care has been taken to avoid unnecessary overhead caused by the high-level programming language. We tested our implementation with the early-stage version of T-Pile written in C. The Java version presented in this Application note is practically as fast as the one written in C language. The Java implementation features high portability and stability of the package. The package contains three programs:

ACorr Computes various autocorrelation and cross-correlation functions, necessary for understanding a system's behavior. ACorr provides also the error analysis utilized by MultiHist.

MultiHist Computes the density of states from the energy observations.

StatPhys Calculates various properties of a molecule as canonical averages. It can read the density of states from a user-specified file or use the output from MultiHist. The other possibility is to use a user-defined partition function as an input.

3 EXAMPLE APPLICATIONS

3.1 Modeling a canonical system

The GB1P peptide (C-terminal beta-hairpin from immunoglobulin binding domain of protein G, PDB code 2gb1) has been extensively studied experimentally (Munoz *et al.*, 1997) and theoretically (Kolinski *et al.*, 1999). Here we used CABS lattice reduced model for studying the folding transition of GB1P. The details of CABS representation and its force field are given elsewhere (Kolinski, 2004). Conformational space was sampled by means of Parallel Tempering, also known as Replica Exchange Monte Carlo. In preliminary runs, we estimated the folding reduced temperature and optimized the set of temperatures for parallel tempering. To estimate statistical errors we performed 12 independent production runs; each for seven replicas, with 1 300 000 observations per replica. T-Pile can process one billion data entries in several minutes on a standard PC.

Figure 1 shows thermodynamic description of the GB1P peptide in the CABS representation: energy as a function of temperature (top-left), heat capacity (top-right). and a 2D $P(E|T)$ conditional distribution (bottom). Highly cooperative, two-state folding is clearly visible.

*To whom correspondence should be addressed.

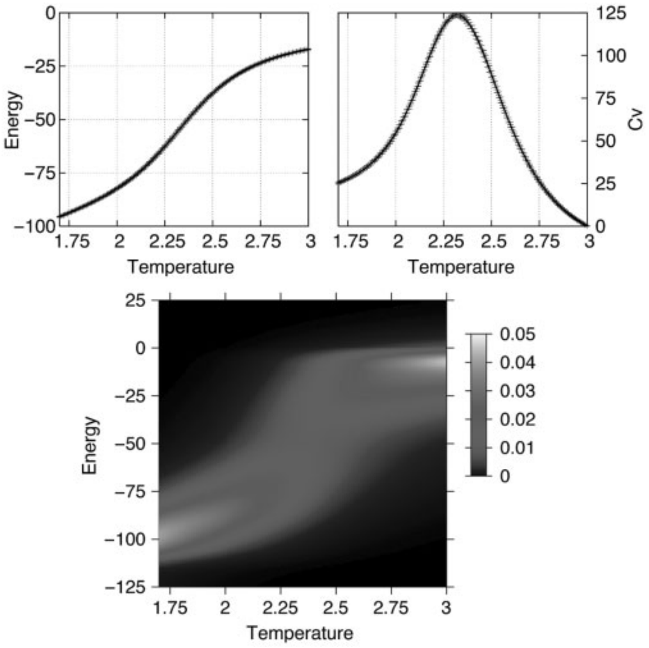

Fig. 1. Mean energy and heat capacity for GB1P peptide (top-left and top-right, respectively) and a 2D histogram showing conditional probability of energy $P(E|T)$ plotted as a function of reduced temperature. Colour version of this figure is available as Supplementary material online.

3.2 Statistical thermodynamics of suboptimal sequence alignments

In a pioneering work, Miyazawa (Miyazawa, 1995) proposed a modification of Needleman–Wunsch algorithm for global sequence alignment in which a single optimal solution is substituted by an ensemble of suboptimal alignments. A free parameter that is introduced into the alignment algorithm plays a role of temperature. Suboptimal alignments obey Boltzmann distribution with the negative score corresponding to ‘energy’. Probability of an alignment with score S is defined as follows:

$$P_T(S) = \frac{1}{\mathcal{Z}(T)}\Omega(S)\exp(S/T) \qquad (1)$$

where $\Omega(S)$ denotes the total number of alignments of score S and $\mathcal{Z}$ is the partition function. It is also possible to estimate a residue–residue alignment probability. For this purpose the partition function $\mathcal{Z}(T)$ for all possible alignments between two sequences, partition function $\mathcal{Z}_{i,j}(T)$ for all possible partial alignments that end at aligned residues i and j, and $\hat{\mathcal{Z}}_{i,j}(T)$—a partition function for all partial alignments that starts with aligned residues i and j need to be computed. Then the probability P_T (i, j) that residues i and j are aligned at temperature T is given by a formula:

$$P_T(i,j) = \frac{1}{\mathcal{Z}(T)}\mathcal{Z}_{i,j}(T)\hat{\mathcal{Z}}_{i,j}(T)\exp(S(i,j)/T) \qquad (2)$$

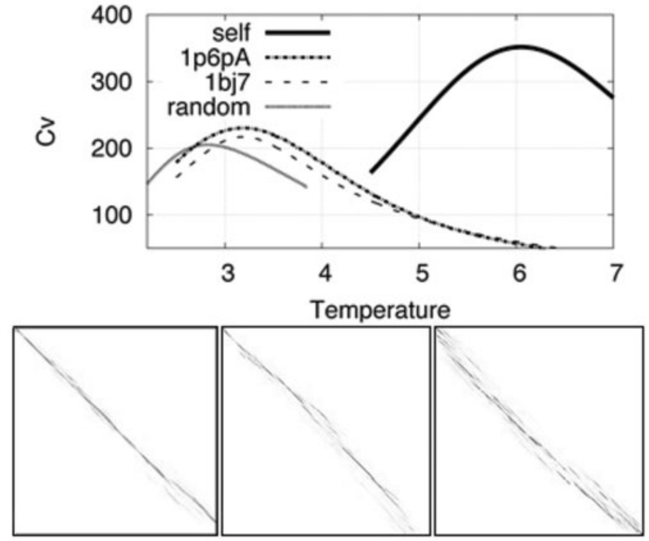

Fig. 2. ‘Heat capacity’ for sequence alignments (upper panel) and probability maps (bottom row) showing how likely residue i from one sequence is aligned with residue j from the other sequence. The three figures are given in the order of decreasing homology from the left to the right: 1b57 − 1p6pA, 1b57 − 1bj7 and 1b57 − 1b57 randomly shuffled. Each probability map was calculated at transition temperature proper for a given pair of sequences: $T = 3.18$, $T = 3.18$ and $T= 2.08$, respectively. Colour version of this figure is available as Supplementary material online.

This algorithm has been implemented in the PRAline tool from BioShell package (Gront and Kolinski, 2006). At $T = 0$, the result of Miyazawa algorithm corresponds to the optimal alignment from Needleman–Wunsch algorithm. With growing temperature, alignments undergo thermal fluctuation and become fuzzy (see Fig. 2). Higher stability of an alignment reflects higher local evolutionary conservation. It is a non-trivial task to find an optimal temperature for the evaluation of the alignment stability. Here, we propose to use the transition temperature as a reference point. The transition temperature corresponds to the peak of heat capacity. To compute Cv for alignments we use StatPhys program from T-Pile package. Figure 2 (top) presents the ‘heat capacity’ for 1b56 protein aligned (dashed line) and with a randomly shuffled sequence of itself. Probability maps (bottom panels) for these alignments (with 1bj7, 1p6pA and random) provide a quantitative measure of the alignments’ stability that could be extremely useful in comparative modeling.

ACKNOWLEDGEMENTS

The computational part of this work was done using the computer cluster at the Computing Center of the Department of Chemistry, University of Warsaw.

Conflict of Interest: none declared.

REFERENCES

Ferrenberg,A.M. and Swendsen,R.H. (1989) Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, **63**, 1195.

Gront,D. and Kolinski,A. (2006) BioShell – a package of tools for structural biology computations. *Bioinformatics*, **22**, 621–622.

Gront,D. *et al.* (2001) A new combination of replica exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. *J. Chem. Phys.*, **115**, 1569–1574.

Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.

Kolinski,A. *et al.* (1999) Dynamics and thermodynamics of beta-hairpin assembly: insights from various simulation techniques. *Biophys. J.*, **77**, 2942–2952.

Kumar,S. *et al.* (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Compu. Chem.*, **13**, 1011–1021.

Miyazawa,S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng*, **8**, 999–1009.

Munoz,V. *et al.* (1997) Folding dynamics and mechanism of [beta]-hairpin formation. *Nature*, **390**, 196–199.