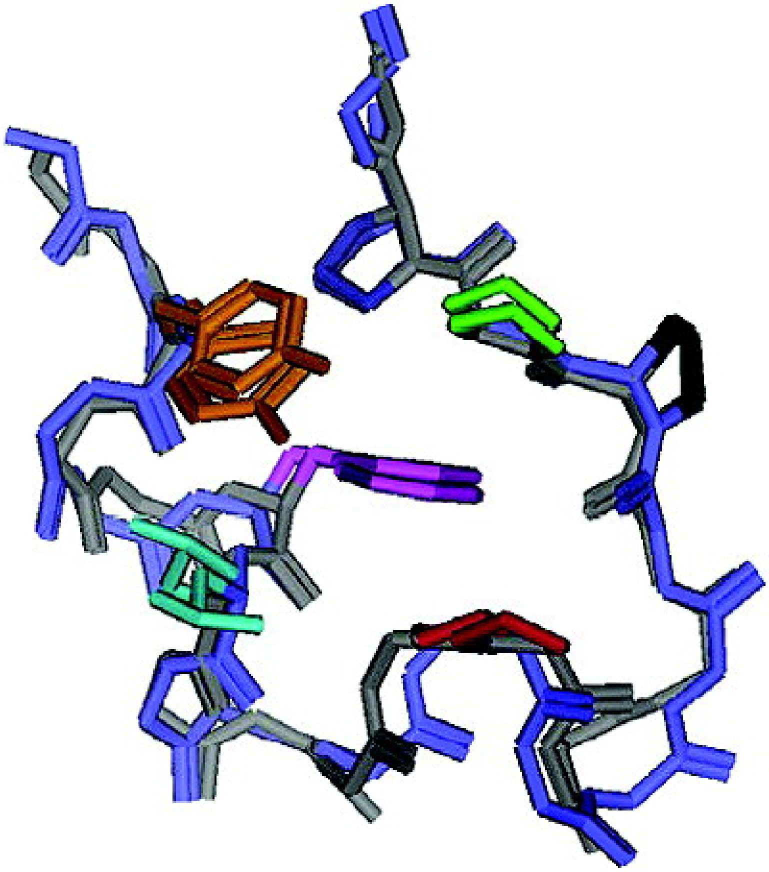
Communication

Fast Protein Structure Prediction Using Monte Carlo Simulations with Modal Moves

Paolo Carnevali, Gergely Tth, Garrick Toubassi, and Siavash N. Meshkat

J. Am. Chem. Soc., **2003**, 125 (47), 14244-14245• DOI: 10.1021/ja036647b • Publication Date (Web): 30 October 2003

Downloaded from http://pubs.acs.org on March 30, 2009



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 4 articles that cite this article, as of the time of this article download

Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

J|A|C|S
COMMUNICATIONS

Published on Web 10/30/2003

Fast Protein Structure Prediction Using Monte Carlo Simulations with Modal Moves

Paolo Carnevali,* Gergely Tóth, Garrick Toubassi, and Siavash N. Meshkat

Protein Mechanics Inc., 280 Hope Street, Mountain View, California 94041

Received June 12, 2003; E-mail: PCarnevali@ProteinMechanics.com

Using normal modes to generate torsion space moves in Monte Carlo simulations of peptides and proteins is not a new idea;[1] nevertheless, despite its power it has not received widespread application.[2] Here, we show that such a "Modal Monte Carlo" approach is an efficient tool for ab initio predictions of small-protein structures. We apply this method to the Trp cage,[3] a 20-residue polypeptide designed to fold rapidly[4] into a structure that includes tertiary contacts, despite its short length. We achieve a high-quality ab initio structure prediction in about 2 orders of magnitude less computation time than state of the art molecular dynamics techniques.[5]

To achieve good sampling efficiency in Monte Carlo simulations, it is necessary to use sets of moves which perturb atomic positions substantially without causing large changes in energy. This is difficult to achieve in folded proteins because of steric interactions. For this reason, when using simple moves such as changes of a single torsion angle, move amplitudes need to be small, and this results in low sampling efficiency. Several types of ad hoc Monte Carlo moves specialized for peptides and proteins have been proposed to overcome this problem, often exploiting a priori knowledge of the folded state.[2,6]

The Modal Monte Carlo approach (often referred to in the literature as the method of scaled collective variables[1]) is a way to generate efficient Monte Carlo moves automatically. We compute normal modes and normalize them in such a way that the vibration energy of each mode, if assumed to be exactly harmonic, is kT. Each Monte Carlo move is then constructed in torsion space as a random linear combination of such temperature-normalized modes. The coefficients of the linear combination can be normalized in such a way that the typical energy change is of the order kT. This ensures an optimal acceptance rate near 50%.

Critical to the validity of this approach is the assumption that most modes are in the near-harmonic regime, which is known to be valid for proteins at ambient temperature.[7] If this were not the case, the energy changes could be much larger than kT. This would result in a decrease of the Monte Carlo acceptance rate with a corresponding degradation of the sampling efficiency.

It is necessary to perform the simulation in torsion space. If Cartesian coordinates are used, normal modes move the atoms along straight lines. This causes large energy increases due to changes in bond lengths and angles, and hence the modes appear to be highly anharmonic. As a result, the Modal Monte Carlo procedure breaks down.

Modal moves are efficient because they result in energy changes of the order kT. Alternatively, the sampling efficiency of this method can be attributed to the fact that all modes are sampled equally, regardless of their frequency. In contrast, molecular dynamics is highly biased in favor of high-frequency modes, which are sampled very generously, at the expense of low-frequency modes, which are sampled at a much slower rate. For this reason the Modal Monte Carlo method has also been described as "Multiclock Simulation".[1] This fact is particularly significant in view of the importance of the lower-frequency modes in protein dynamics.[8] A study of the sampling efficiency of the Modal Monte Carlo method found it to be better by a factor of 50 compared to using straightforward moves.[1]

One problem with the procedure as originally proposed[1] is the presence of large numbers of exponential modes, which are dealt with in an ad hoc fashion. We avoid this problem altogether by performing a local energy minimization before each recomputation of the normal modes. The simulation then continues from the pre-minimization conformation.

It has long been known[9] that a straightforward application in torsion space of a force field developed for Cartesian coordinates is problematic. For this reason we have used a force field projection scheme similar to one recently proposed.[10] Such a scheme includes averaging of torsion terms to account for changes in bond lengths and angles; softening of van der Waals interactions for 1−5 and 1−6 atom pairs; and optimizing atom positions within each rigid body comprising the torsion space model. This systematic procedure improves substantially the quality of the force field when working in torsion coordinates and proved to be an essential ingredient for high-quality structure predictions.

We chose the Trp cage[3,11] (TC5b sequence) as the test application of Modal Monte Carlo, since it is an ideal model system for folding simulations.[3] We ran a set of Modal Monte Carlo simulations using the unmodified Amber 94[12] all-atoms force field with water modeled using the generalized Born/surface area implicit solvent formulation.[13] The calculations were performed using the software package Imagiro, developed at our organization. All runs started from an extended conformation of the Trp cage, and we used a variety of annealing schedules (see Supporting Information for details). The lowest-energy structure was obtained with a run at a constant temperature of 500 K. This run consisted of 25×10^6 attempted Monte Carlo steps, with the normal modes recomputed every 5000 attempted steps. The lowest-energy structure was obtained after 19×10^6 attempted Monte Carlo steps, which required approximately 4.5 days of computation on a single processor of an AMD Athlon MP 2400+ computer. Even after about 1 day of computation time most of the sampled conformations were already native. The prediction of the structure of the Trp cage using molecular dynamics[5] required a total computing time equivalent to about one year on a single computer. The Modal Monte Carlo approach is a substantial improvement in sampling efficiency which could turn ab initio predictions of small-protein structures into a matter of routine.

A comparison of the conformation of the computed structure with the average experimental structure from NMR data[3] is shown in Figure 1. The average of the root-mean-square displacements (rms) over the 38 NMR structures is 1.3 ± 0.2 Å when considering only C_α atoms except for the two terminal ones, and 1.7 ± 0.2 Å when considering all heavy atoms, except for the two terminal

10.1021/ja036647b CCC: $25.00

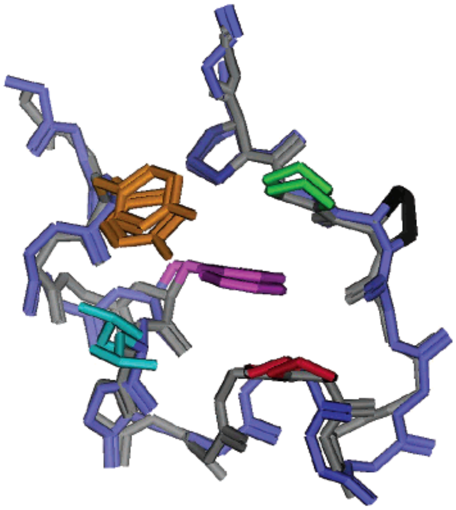


Figure 1. Comparison of the computed lowest-energy structure (blue backbone) with the average NMR structure (gray backbone). Selected side chains are color-coded as follows: Tyr22 (orange), Trp25 (magenta), Leu26 (cyan), Pro31 (dark red), Pro36 (black), and Pro38 (blue). Not all side chains are shown. This structure was obtained in 19 × 10^6 Monte Carlo steps.

residues, and for the side chains of Leu2, Lys8, and Arg16, which were found to be highly flexible in the NMR data. The uncertainties shown are standard deviations. These rms displacements compare well with the corresponding values 1.0 and 1.4 Å obtained with molecular dynamics simulations before refinement in explicit water.[5] The slightly lower quality of our results may be due to the fact that the molecular dynamics simulations were done using a modified version of the Amber 99 force field,[14] while we used an unmodified version of the older Amber 94 force field[12] (projected into torsion space as described above).

The lowest-energy structure contains two α-helices between Leu21 and Lys27 and between Gly30 and Ser33, and a polyproline II helix between Pro36 and Pro38. The Trp cage motif is well matched in the lowest-energy structure. As in the native structure, pyrolidone rings of Pro are located on both faces of the indole ring of Trp25 forming CH−π interaction.[15] In addition, the indole ring forms (*i*) aromatic−aromatic interaction[16] with the hydroxyphenyl ring of Tyr22, (*ii*) aromatic−backbone amide interaction[17] with NH of Gly30, and (*iii*) CH−π interactions with CHα of Leu26 and Gly30. The hydrogen bond between the backbone amide of Gly30 NH and Trp25 carbonyl was found. The hydrogen bond between indole NHϵ1 hydrogen and Asp 35 carbonyl was not found, although these groups are in close proximity. The salt bridge between the side chain of Arg35 and the γ-carboxyl group of Asp28 was found as in molecular dynamics simulations.[5]

NMR ring shifts (σ_{ring}) were calculated from the NMR ensemble and the lowest-energy structure using the Total program.[18] Highly stereospecific σ_{ring}'s were matched well, such as Gly30 NH −0.76 ± 0.05 ppm vs −0.84 ppm and Gly30 CHα −2.76 ± 0.29 ppm/−0.94 ± 0.19 ppm vs −2.39 ppm/−0.58 ppm for the NMR structures and our lowest-energy structure, respectively. The correlation between the calculated σ_{ring} of the NMR ensemble and the lowest-energy structure is 0.89. The rms deviation between the two data sets is 0.16 ppm. The proton with the highest σ_{ring} from Pro31, Pro37, and Pro38 was excluded from the dataset.

Other researchers have also achieved successful structure predictions for systems similar in size to the Trp cage using a different Monte Carlo approach in torsion space.[6] However, these results cannot be considered fully ab initio predictions since they use statistical information from the Protein Data Bank database in the construction of Monte Carlo moves.

The results presented here show that the Modal Monte Carlo approach can be used for fast, high-quality ab initio predictions of small-protein structures, and constitutes a step toward turning ab initio structure prediction for such systems into a routine procedure. Studies of other small proteins using more recent force fields are in progress and will be reported on separately. Application to larger proteins will require a study of the scaling of the method with problem size, which is in progress. We are also investigating applicability to nucleotide structures and polymers in general, and to protein−ligand and protein−protein complexes.

Acknowledgment. The software package Imagiro used in this work was developed by Michael Sherman, Dan Rosenthal, Daniel Filip, and Erin Catto, in collaboration with us. Michael Levitt, Jay Ponder, and Michael Sherman provided valuable input on early drafts of this paper.

Supporting Information Available: Details of the runs performed (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Noguti, T.; Gō, N. *Biopolymers* **1985**, *24*, 527−546.
(2) Scheraga, H. A.; Pillardy, J.; Liwo, A.; Lee, J.; Czaplewski, C.; Ripoll, D. R.; Wedemeyer, W. J.; Arnautova, Y. A. *J. Comput. Chem.* **2002**, *23*, 28−34.
(3) (a) Neidigh, J. W.; Fesynmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425−430. (b) Gellman, S. H.; Woolfson, D. N. *Nat. Struct. Biol.* **2002**, *9*, 408−410.
(4) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952−12953.
(5) (a) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259. (b) University of Florida Press release, October 17, 2002.
(6) (a) Abagyan, R. A.; Totrov, M. *J. Comput. Phys.* **1999**, *151*, 402−421. (b) Totrov, M.; Abagyan, R. A. *Biopolymers (Pept. Sci.)* **2001**, *60*, 124−133.
(7) Brooks, B.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1983**, *80*, 6571−6575.
(8) Levitt, M.; Sander, C.; Stern, P. S. *J. Mol. Biol.* **1985**, *181*, 423−447.
(9) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59−107.
(10) Katritch, V.; Totrov, M.; Abagyan, R. *J. Comput. Chem.* **2002**, *24*, 254−265.
(11) Snow, C. D.; Zagrovic, B.; Pande, V. S. *J. Am. Chem. Soc.* **2002**, *124*, 14548−14549.
(12) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(13) Qiu, D.; Shenkin, P.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.
(14) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(15) Tóth, G.; Murphy, R. F.; Lovas, S. *Protein Eng.* **2001**, *14*, 543−547.
(16) Burley, S. K.; Petsko, G. A. *Science* **1985**, *229*, 23−28.
(17) Tóth, G.; Watts, C. R.; Murphy, R. F.; Lovas, S. *Proteins: Struct. Funct. Genet.* **2001**, *43*, 373−381.
(18) Williamson, P. W.; Asakura, T. *J. Magn. Reson. B* **1993**, *101*, 67−71.

JA036647B