Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9886–9889, October 1995
Biophysics

Enhanced conformational sampling in Monte Carlo simulations of proteins: Application to a constrained peptide

(entropy-sampling Monte Carlo/scaled-collective-variable Monte Carlo/cell-adhesive Arg-Gly-Asp sequence)

AKINORI KIDERA†

Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

Communicated by Harold A. Scheraga, Cornell University, Ithaca, NY, June 27, 1995

ABSTRACT **A Monte Carlo simulation method for globular proteins, called extended-scaled-collective-variable (ESCV) Monte Carlo, is proposed. This method combines two Monte Carlo algorithms known as entropy-sampling and scaled-collective-variable algorithms. Entropy-sampling Monte Carlo is able to sample a large configurational space even in a disordered system that has a large number of potential barriers. In contrast, scaled-collective-variable Monte Carlo provides an efficient sampling for a system whose dynamics is highly cooperative. Because a globular protein is a disordered system whose dynamics is characterized by collective motions, a combination of these two algorithms could provide an optimal Monte Carlo simulation for a globular protein. As a test case, we have carried out an ESCV Monte Carlo simulation for a cell adhesive Arg-Gly-Asp-containing peptide, Lys-Arg-Cys-Arg-Gly-Asp-Cys-Met-Asp, and determined the conformational distribution at 300 K. The peptide contains a disulfide bridge between the two cysteine residues. This bond mimics the strong geometrical constraints that result from a protein's globular nature and give rise to highly cooperative dynamics. Computation results show that the ESCV Monte Carlo was not trapped at any local minimum and that the canonical distribution was correctly determined.**

Globular proteins have extremely rugged potential surfaces, which are often characterized as the multiple minima (1) or the conformational substates (2). In such a disordered system, adequate conformational sampling is difficult to achieve, particularly at low temperatures because of the potential barriers that surround low-energy regions. As a result, traditional simulation methods often become trapped in a local energy minimum close to the starting structure.

Recently, a multicanonical (entropy sampling) Monte Carlo algorithm‡ has been proposed as a way of alleviating the sampling difficulty typically observed in disordered systems (3–6). This method samples all energy levels equally by introducing an energy-dependent weight function to the Metropolis Monte Carlo scheme (7). Conceptually, it corresponds to a simulation in which high (low)-energy regions are sampled at high (low) temperatures. In this manner, potential barriers can be overcome at high temperatures, whereas the details of the low-energy potential surface can be traced at low temperatures. The canonical distribution at any temperature can be calculated from the sampled ensemble by the reweighting technique (8). This Monte Carlo algorithm has been used to determine the low-temperature ensembles of various disordered systems (3, 4, 9–11), including a small linear peptide (12) and lattice proteins (13–15).

However, in addition to the sampling problem, we have to find a solution for the problem of low efficiency, or low acceptance ratio, commonly associated with the application of Monte Carlo simulation to a globular protein. It is difficult to change one torsion angle without generating unrealistic interatomic distances within a molecule. This problem is caused by the highly cooperative nature of a globular protein, which results from the strong geometrical constraints such as the tightly packed globular shape and the unique local conformations (16). Noguti and Go (17) introduced a set of collective variables for Monte Carlo simulation, which are derived from a Hessian matrix of the potential energy and contain information about protein dynamics. Their scaled-collective-variable (SCV) algorithm has allowed efficient Monte Carlo simulations of globular proteins (18, 19).

A combination of these two algorithms, called extended-SCV (ESCV) Monte Carlo, might therefore be a promising way for improving the conformational sampling of a globular protein. In this study, as a simple test case, we have carried out an ESCV Monte Carlo simulation of a nine-residue peptide, Lys-Arg-Cys-Arg-Gly-Asp-Cys-Met-Asp, and have determined the conformational distribution at 300 K. We chose this peptide for two reasons. First, this peptide contains the cell-adhesive Arg-Gly-Asp sequence and derives from a disulfide mutant of echistatin, a potent integrin antagonist (T. Yamada and A.K., unpublished results). As the conformation of the functional Arg-Gly-Asp sequence is an important issue in understanding the molecular recognition of integrins, this represents an interesting application of the method. Second, the peptide has a strong geometrical constraint imposed by a disulfide bond between Cys-3 and Cys-7. This constraint ideally mimics effects in globular proteins that result in the low efficiency of the traditional Monte Carlo method.

In the following section, we present the method of the ESCV Monte Carlo and then show the results of the simulation.

THEORY: ESCV MONTE CARLO

We consider Metropolis Monte Carlo (7) of a globular protein in the dihedral angle space. A trial move from conformation m to n (from a set of dihedral angles $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_n$) is accepted with a probability, min(1, $\alpha_{nm}\rho_n/\alpha_{mn}\rho_m$), where α_{mn} is *a priori* transition probability ($m \rightarrow n$) and ρ_m is the Boltzmann probability of conformation m *(20)*.

SCV Monte Carlo. One of the most characteristic features in protein dynamics is that each degree of freedom, the dihedral angles in this case, is highly correlated with one another (16). Trial steps should be chosen so as to reflect such a collective nature of the protein dynamics. In the SCV Monte Carlo (17), *a priori* transition probability, α_{mn}, is given by considering the correlation of motion in the form of a multivariate normal distribution,

$$\alpha_{mn} = N \exp\left(-\frac{1}{2}\boldsymbol{\theta}_{mn}^{t}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_{mn}\right), \qquad [1]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SCV, scaled collective variable; ESCV, extended SCV.

[†] Present address: Biomolecular Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan.

[‡] There are two algorithms in this context, multicanonical (3) and entropy sampling (4), whose comparisons are found in refs. 5 and 6.

where $\boldsymbol{\theta}_{mn}$ ($= \boldsymbol{\theta}_n - \boldsymbol{\theta}_m$) is a displacement vector in the dihedral angle space; superscript means transpose; $\boldsymbol{\Sigma}$ is a covariance matrix of $\boldsymbol{\theta}_{mn}$, representing the correlation of dihedral angles; and N is a normalization constant. The covariance, $\boldsymbol{\Sigma}$, can be estimated by a harmonic approximation,

$$\boldsymbol{\Sigma} \approx kT\mathbf{F}^{-1}, \qquad [2]$$

where T is temperature and $\mathbf{F}$ is a Hessian matrix [$= \{\partial^2 E/\partial\theta_i\partial\theta_j\}$] of the potential energy E at conformation m. With use of the eigenvalues, λ_k, and the eigenvectors, $\boldsymbol{\omega}_k$, of $\mathbf{F}$ ($\mathbf{F}\boldsymbol{\omega}_k = \lambda_k\boldsymbol{\omega}_k$; $k = 1, \ldots, n$, where n is the total degrees of freedom), Eq. 1 is rewritten in the scaled variable form,

$$\alpha_{mn} = N\prod_k \exp\left(-\frac{1}{2kT}\phi_k^2\left|\lambda_k\right|\right), \qquad [3]$$

where ϕ_k is the kth collective variable given by $\boldsymbol{\omega}_k^t\boldsymbol{\theta}_{mn}$, which is scaled by $|\lambda_k|^{-1/2}[\sim(\langle\phi_k^2\rangle/\mathrm{kT})^{1/2}]$. The absolute symbol of λ_k is due to the fact that $\mathbf{F}$ is not necessarily positive definite at conformation m. Trial moves proceed along the SCVs, ϕ_k.

In the actual implementation, ϕ_k and λ_k are updated every 100 steps or less frequently. Hence, in most steps, the same set of the collective variables is used for the reverse step, $n \rightarrow m$, so that $\alpha_{nm}/\alpha_{mn} = 1$. Only at the steps where the collective variables are updated did α_{nm}/α_{mn} deviate from unity. It has been confirmed that these breaks of microscopic reversibility do not cause any significant bias to the sampled ensemble. Therefore, the Metropolis criterion can be $\min(1, \rho_n/\rho_m)$.

Entropy-Sampling Monte Carlo. Our purpose for the Monte Carlo simulation is to determine the probability distribution of a wide range of conformations at an ordinary temperature ($\approx$300 K). For this purpose, the simulation has to sample both low-energy and high-energy conformations. The former should be significant in the probability distribution at 300 K, and the latter is required to cross over potential barriers. Such a simulation can be done on a modified potential surface, giving a flat energy distribution. Here, we follow mostly the algorithm of entropy-sampling Monte Carlo (4, 6).

The entropy-sampling method modifies the potential surface by introducing an energy-dependent weight function, $w(E)$, to the Metropolis scheme as

$$\min(1, \rho_n/\rho_m) = \min(1, \exp[-w(E_n) + w(E_m)]) \qquad [4]$$

for the conventional form, $\rho_n/\rho_m = \exp(-E_n/kT + E_m/kT)$. The function, $w(E)$, is chosen in such a way that a Monte Carlo simulation with the criterion of Eq. 4 would result in a flat energy distribution, P_w,

$$P_w(E) = Z_w^{-1}n(E)e^{-w(E)} = \text{constant}, \qquad [5]$$

where $\mathrm{n}(E)$ is the spectral density and $Z_w = \Sigma_E\, ne^{-w}$. Even from such an artificial ensemble of conformations, the canonical distribution at any temperature T, $P_B(E;T)$, can be correctly recovered by the reweighting formula (8),

$$P_B(E;T) = \mathrm{Z}^{-1}\mathrm{n(E)e}^{-\mathrm{E/kT}} = \mathrm{Z}^{-1}Z_w e^{w(E)-E/kT}P_w(E), \qquad [6]$$

where $Z = \Sigma_E ne^{-E/kT}$.

In practice, the function, $w(E)$, is determined with the help of a preliminary canonical Monte Carlo run by the SCV method at a sufficiently high temperature T^*,

$$w(E) = \ln n(E) = E/kT^* + \ln P_B(E;T^*), \qquad [7]$$

where the energy-independent terms are neglected. It is possible to refine the functional form of $w(E)$ iteratively by using the relationship derived from Eq. 5—i.e., $w^{i+1}(E) = w^i(E) + \ln P_w^i(E)$.

In summary, the ESCV Monte Carlo is performed by employing *a priori* transition probability, α_{mn}, given by the SCV algorithm (Eq. 3) and the term, ρ_n/ρ_m, evaluated by the entropy-sampling algorithm (Eq. 4).

COMPUTATION

The simulation system is a nine-residue peptide, Lys-Arg-Cys-Arg-Gly-Asp-Cys-Met-Asp, in which Cys-3 and Cys-7 form a disulfide bond. In addition to the S–S bridge, a weak distance restraint (3 kcal/mol) is imposed between the C_α atoms of the two terminal residues (Lys-1 and Asp-9) to maintain the distance observed in the NMR data of echistatin (21). The force field parameters used are those of ECEPP/2 (22), in which all ionizable groups are in the neutral state. No solvent molecule is considered in the simulation.

For the SCV Monte Carlo, the Hessian matrix in Eq. 2 ($\mathbf{F}$) is updated every 100 Monte Carlo steps by using FEDER (23), a program for fast Hessian calculation. To prevent the scaling factor in Eq. 3, $|\lambda_k|^{-1/2}$, from becoming too large, we used λ_{limit} [$= 3$ kcal/(mol·rad^2)] for $|\lambda_k|$ when $|\lambda_k| < \lambda_{\text{limit}}$. The step length was chosen so that the acceptance ratio would be about 0.5.

The weight function, $w(E)$, for an ESCV Monte Carlo was determined in the following manner. A preliminary SCV Monte Carlo run at T* ($= 1000$ K) accumulated up to 10^6 steps. Such a high temperature was necessary to cover all possible conformations without suffering from the multiple-minima problem. The resultant probability distribution, $P_B(E;T^*)$, with a bin size of 1 kcal/mol, was fitted and extrapolated by a fourth-order polynomial (Fig. 1). To avoid the oversampling of very low- and high-energy regions, the extrapolation by the polynomial was continued smoothly by a linear function for the regions of $E \leq -80$ and $E \geq 0$ (kcal/mol). With the extrapolated canonical distribution, $P_B(E;T^*)$, the weight function, $w(E)$, was determined by Eq. 7. A refinement of $w(E)$ by the iteration formula was not performed for the following reason. Once the simulation begins to cover the low-energy region, this system requires a very long Monte Carlo run to reach equilibrium. When considering an application of the ESCV method to globular proteins, it is desirable to have a method of giving a good estimate of $w(E)$ that does not necessarily require a refinement.

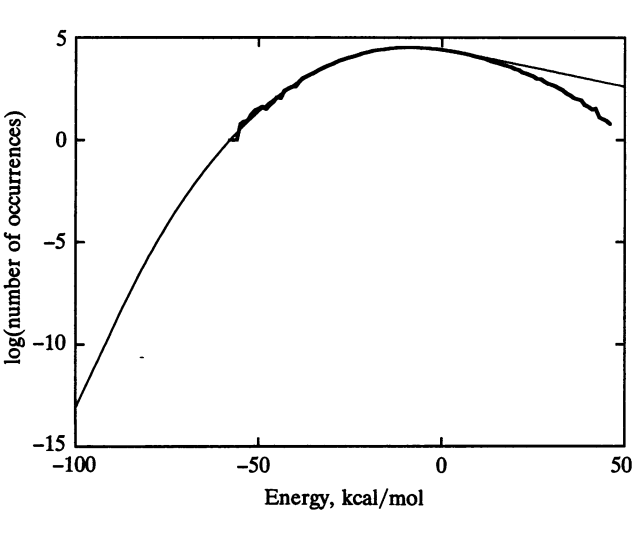

FIG. 1. Determination of the weight function, $w(E)$, of Eq. 4. The thick curve is the energy distribution (bin size of 1 kcal/mol) of the preliminary canonical SCV Monte Carlo of 10^6 steps at 1000 K. The thin curve is the fourth-order polynomial fitted to the thick curve, from which $w(E)$ was determined by Eq. 7 (see text for details).

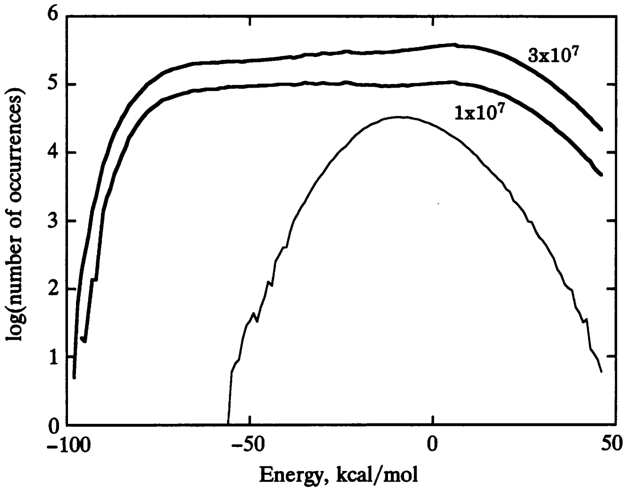

FIG. 2. Energy distribution given by the ESCV Monte Carlo simulation of the peptide. The thick curves are those after the accumulation of 1 × 10^7 steps and 3 × 10^7 steps, whereas the thin curve is the energy distribution of the canonical SCV Monte Carlo at 1000 K.

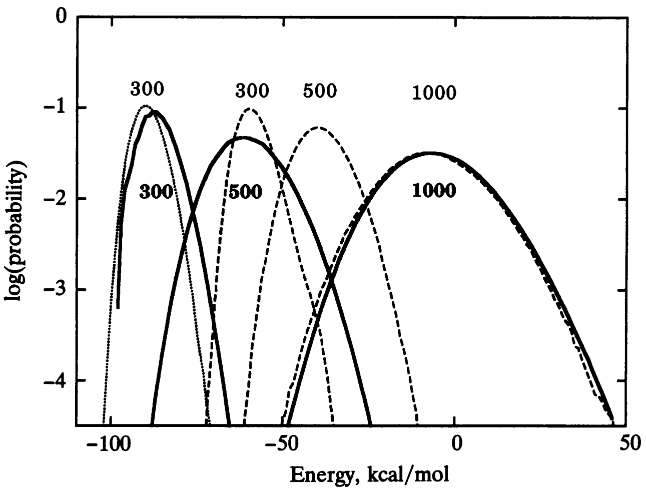

FIG. 3. Energy distribution of the canonical ensemble at 300, 500, and 1000 K calculated by reweighting the sampled ensemble (the thick solid curves). The broken curves are those of the canonical SCV Monte Carlo runs of 5 × 10^6 steps, started from the conformation of step 2 × 10^7 of the ESCV Monte Carlo run. The dotted curve is the same as the broken curve but started at the minimum energy conformation found in the ESCV Monte Carlo.

An ESCV Monte Carlo simulation using $w(E)$ of Fig. 1 was carried out to accumulate 3 × 10^7 Monte Carlo steps. All computations were done on a Fujitsu VP2600 (Tokyo).

RESULTS AND DISCUSSION

Monte Carlo Simulations. An ESCV Monte Carlo simulation with the weight function $w(E)$ shown in Fig. 1 resulted in a flat energy distribution (Fig. 2). The flat distribution, which had already converged after 1 × 10^7 Monte Carlo steps, confirmed that the sampling had covered all energy levels and was thus suitable for evaluating the canonical distribution.

The reweighting operation of Eq. **6** to the sampled ensemble yields the canonical distributions at various temperatures shown in Fig. 3. This figure also shows the energy distributions derived from canonical Monte Carlo runs (i.e., SCV Monte Carlo simulations using the conventional canonical criterion; non-SCV runs did not succeed because most of moves de-

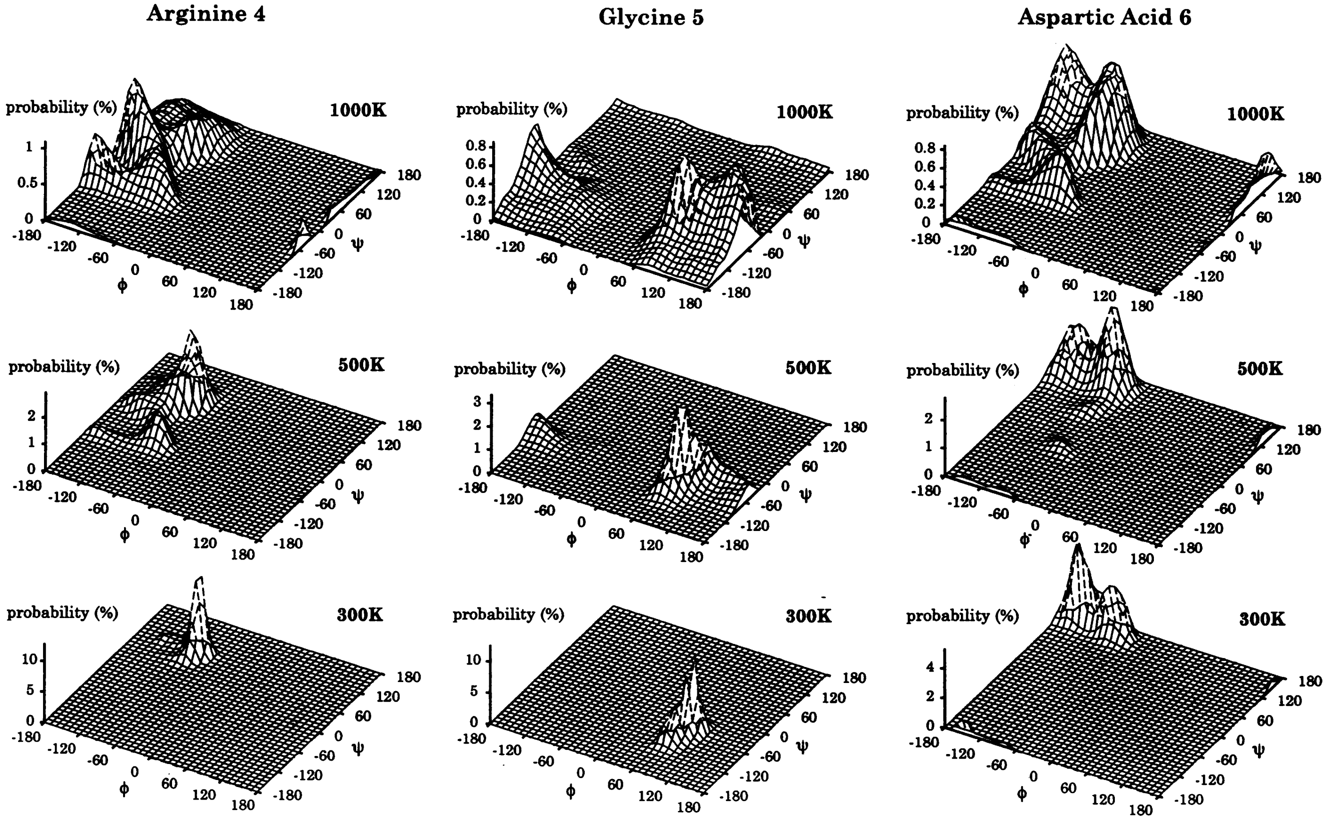

FIG. 4. The probability distributions of the main-chain dihedral angles, ϕ and ψ, for the Arg-Gly-Asp residues at 300, 500, and 1000 K, calculated by reweighting the sampled ensemble.

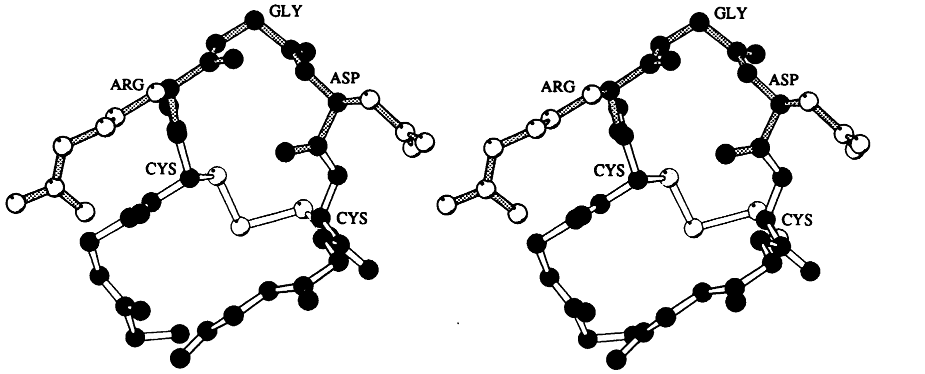

FIG. 5. Stereo Molscript drawing (24) of a representative structure of the peptide at 300 K found in the ESCV Monte Carlo simulation. Only the side-chain atoms of the Arg-Gly-Asp and the cysteine residues are given explicitly.

stroyed the disulfide bond). At 1000 K, the canonical simulation gives almost the same distribution as the ESCV simulation. However, at lower temperatures, large discrepancies are observed, with canonical simulations at 300 and 500 K becoming trapped in high-energy conformations. This leads to distributions with energies higher by 20–30 kcal/mol. However, when a canonical simulation is started from a low-energy conformation found by the ESCV simulation, good agreement was obtained. This shows that canonical Monte Carlo cannot surmount potential barriers and is thus dependent on the initial conformation.

In conclusion, the ESCV Monte Carlo method is capable of correctly determining a probability distribution of a constrained peptide at 300 K. The results described above indicate that the ESCV Monte Carlo can be a promising algorithm for improving the conformational sampling of disordered and cooperative systems such as globular proteins.

RGD Conformations. From a biological point of view, the cell-adhesive Arg-Gly-Asp sequence of our test peptide is the most important. Fig. 4 shows the probability distributions of the main chain ϕ–ψ angles of the Arg-Gly-Asp residues at 1000, 500, and 300 K. At 1000 K, Arg-4 and Asp-6 show ordinary alanine-like distributions, whereas in Gly-5, conformations with $\psi > 0$ are strongly suppressed by the geometrical constraint of the disulfide bond. At lower temperatures, the conformational distributions become localized and attain an almost single conformation at 300 K. The peaks of the distribution locate at $(\phi, \psi) = (-70, 90)$, $(90, -50)$, and $(-90$ or $-150, 160)$ for Arg, Gly and Asp, respectively. A representative structure of the peptide is displayed in Fig. 5.

Recently, we showed (25) that the Arg-Gly-Asp conformations can be divided mainly into two structural classes in terms of the distance d between the C_β atoms of arginine and aspartic acid residues and of the conformation of glycine residue: in class 1, $d = 9.0$–9.5 Å and glycine conformation = E, F, E^*, or F^* [flavoridin (26) and foot-and-mouth disease virus (27)]; in class 2, $d = 7.5$–8.5 Å and glycine conformation = C^* or D^* [tenascin (28), decorsin (29), and an Arg-Gly-Asp-containing mutant of human lysozyme (30)]. Here, the conformations are designated according to Zimmerman *et al.* (31), and the name lists in brackets are the cell-adhesive proteins whose Arg-Gly-Asp x-ray (or NMR) structures are well-defined and classified into the respective classes.

By these criteria, the conformation of Fig. 5 is categorized as class 2. The Monte Carlo simulation shows that the geometrical constraint of the disulfide bond prohibits the conformation of class 1 (Gly-5 conformation with $\psi > 0$). Experimental data indicate that the echistatin mutant, containing this peptide as a fragment, is a more potent integrin antagonist than wild-type echistatin, which has no disulfide constraint in the Arg-Gly-Asp region (T. Yamada & A.K., unpublished results). These facts suggest that the active conformation of the Arg-Gly-Asp sequence would be class 2.

I thank Dr. M. B. Swindells for helpful discussions.

1. Vásquez, M., Némethy, G. & Scheraga, H. A. (1994) *Chem. Rev.* **94**, 2183–2239.
2. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254**, 1598–1603.
3. Berg, B. A. & Neuhaus, T. (1991) *Phys. Lett.* **B267**, 249–253.
4. Lee, J. (1993) *Phys. Rev. Lett.* **71**, 211–214.
5. Berg, B. A., Hansmann, U. H. E. & Okamoto, Y. (1995) *J. Phys. Chem.* **99**, 2236–2237.
6. Hao, M.-H. & Scheraga, H. A. (1995) *J. Phys. Chem.* **99**, 2238.
7. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21**, 1087–1092.
8. Ferrenberg, A. M. & Swendsen, R. H. (1988) *Phys. Rev. Lett.* **61**, 2635–2638.
9. Berg, B. A. & Celik, T. (1992) *Phys. Rev. Lett.* **69**, 2292–2295.
10. Berg, B. A., Hansmann, U. & Neuhaus, T. (1993) *Phys. Rev.* **B47**, 497–500.
11. Berg, B. A., Celik, T. & Hansmann, U. (1993) *Europhys. Lett.* **22**, 63–68.
12. Hansmann, U. & Okamoto, Y. (1993) *J. Comput. Chem.* **14**, 1333–1338.
13. Hao, M.-H. & Scheraga, H. A. (1994) *J. Phys. Chem.* **98**, 4940–4948.
14. Hao, M.-H. & Scheraga, H. A. (1994) *J. Phys. Chem.* **98**, 9882–9893.
15. Hao, M.-H. & Scheraga, H. A. (1994) *J. Chem. Phys.* **102**, 1334–1348.
16. Hayward, S. & Go, N. (1995) *Annu. Rev. Phys. Chem.*, in press.
17. Noguti, T. & Go, N. (1985) *Biopolymers* **24**, 527–546.
18. Go, N. & Noguti, T. (1989) *Chem. Scripta* **29A**, 151–164.
19. Horiuchi, T. & Go, N. (1991) *Proteins* **10**, 106–116.
20. Allen, M. P. & Tildesley, D. J. (1987) *Computer Simulation of Liquids* (Oxford Univ. Press, Oxford).
21. Saudek, V., Atkinson, R. A. & Pelton, J. T. (1991) *Biochemistry* **30**, 7369–7372.
22. Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983) *J. Phys. Chem.* **87**, 1883–1887.
23. Wako, H. & Go, N. (1987) *J. Comput. Chem.* **8**, 625–635.
24. Kraulis, J. P. (1991) *J. Appl. Crystallogr.* **24**, 946–950.
25. Yamada, T., Uyeda, A., Kidera, A. & Kikuchi, M. (1994) *Biochemistry* **33**, 11678–11683.
26. Senn, H. & Klaus, W. (1993) *J. Mol. Biol.* **232**, 907–925.
27. Logan, D., Abu-Ghazaleh, R., Blakemore, W., Curry, S., Jackson, T., King, A., Lea, S., Lewis, R., Newman, J., Parry, N., Rowlands, D., Stuart, D. & Fry, E. (1993) *Nature (London)* **362**, 566–568.
28. Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992) *Science* **258**, 987–991.
29. Krezel, A. M., Wagner, G., Seymour-Ulmer, J. & Lazarus, R. A. (1994) *Science* **264**, 1944–1947.
30. Yamada, T., Song, H., Inaka, K., Shimada, Y., Kikuchi, M. & Matsushima, M. (1995) *J. Biol. Chem.* **270**, 5687–5690.
31. Zimmerman, S. S., Pottle, M., Némethy, G. & Scheraga, H. A. (1977) *Macromolecules* **10**, 1–9.