# Computer Simulation of the Entropy of Polypeptides Using the Local States Method: Application to Cyclo-(Ala-Pro-D-Phe)$_2$ in Vacuum and in the Crystal

**Hagai Meirovitch,*,[†],[‡] David H. Kitson,[†] and Arnold T. Hagler[†]**

*Contribution from Biosym Technologies, Inc., 9685 Scranton Road, San Diego, California 92121, and Supercomputer Computations Research Institute, Florida State University, Tallahasse, Florida 32306.* *Received December 3, 1991*

**Abstract:** The local states (LS) method is an approximate technique proposed by Meirovitch (*Chem. Phys. Lett.* **1977**, *45*, 38) for estimating the entropy from a sample of conformations. The method is further developed and extended here to molecular dynamics samples of the cyclic peptide cyclo-(Ala-Pro-D-Phe)$_2$ in vacuum and in the crystal environment. This is the first time the LS method has been applied to a peptide with side chains which is described by flexible geometry. The method enables one to obtain an approximation $P_i$ of the sampling probability of conformation $i$ where $P_i$ is expressed as a product of transition probabilities, which relate a dihedral or a valence angle to a number of preceding angles in the chain. This set of angles is called a local state. The values of $P_i$ define approximations for the entropy $S$ ($S \sim \ln P_i$) which together with the energy lead to upper and lower bounds for the free energy. The LS method is general; i.e., it can be applied to samples of any conformational state (e.g., a random coil) and is not restricted to a molecule undergoing small harmonic conformational fluctuations. Thus, the relative stability of states of considerable structural difference can be obtained from the corresponding free energies. In this work we investigate some effects of environment on cyclo-(Ala-Pro-D-Phe)$_2$ and, in particular, calculate the reduction in entropy, $\Delta S$, in going from vacuum to the crystal. We find that $T\Delta S = T[S(\text{vacuum}) - S(\text{crystal})] = 9.4 \pm 0.8$ kcal/mol where $T \sim 300$ K is the absolute temperature. Calculation of $\Delta S$ is important in biological processes such as the binding of a peptide to a receptor, which involves a change of environment. We argue that under certain conditions the method is expected to be efficient even for large proteins.

## I. Introduction

Polypeptides and proteins are known to have a large number of metastable states, i.e., states that correspond to relatively low Helmholtz free energy, $F$. An important theoretical goal is to calculate this free energy, which constitutes the criterion of stability. There are a variety of applications for which the ability to calculate free energy is useful. For example, in studying protein-ligand interactions, one wants to know the free energy changes involved in the binding process, or the relative free energy of binding of different ligands (see refs 1-3 and references cited therein). It is also important to be able to compare the relative free energies of different conformational states of a peptide (such as an $\alpha$-helical or a $\beta$-turn structure) in different environments (in vacuo, in solution, in an enzyme active site, etc.). However, with the commonly used computer simulation techniques, Metropolis Monte Carlo[4] and molecular dynamics,[5,6] calculation of the entropy, $S$ (and hence $F$), is difficult. This stems from the fact that these methods are of a dynamical type (i.e., one starts from some conformation which evolves in time), and they therefore do not provide the *value* of the sampling probability, $P$, of a conformation, that leads to the entropy, $S$ ($S \sim \ln P$). Thus, in many studies, the energy, $E$ (which can be obtained easily), rather than $F$, has been adopted as an approximate criterion of stability.

Several approximate methods for calculating the entropy of macromolecules have been proposed. Gō and Scheraga[7,8] developed a method which, in principle, is based on a normal coordinates analysis (using classical statistical mechanics) for calculating the conformational entropy of macromolecules undergoing small (i.e., harmonic) fluctuations around their stable state (e.g., the $\alpha$-helical state of a polypeptide). The method was applied to several polypeptides[9,10] modeled by rigid geometry, i.e., fixed bond lengths and bond angles. They also calculated the entropy of a polypeptide in its random-coil state[7] at the $\theta$ point,[9-12] i.e., by neglecting most of the excluded volume effect. Hagler et al.[13] introduced the Einstein harmonic oscillator approximation to the calculation of the entropy of different conformational states of peptides as well as to the effects of residue substitutions in different positions. They also demonstrated the importance of flexibility, i.e., of allowing for the relaxation of bond lengths and angles in determining conformational stability. Karplus and Kushick[14] have proposed to calculate the covariances of the internal coordinates directly from a molecular dynamics or a Monte Carlo simulation (the quasiharmonic approximation[15]) and applied their method to butane and decaglycine (see also refs 16-18). These approximate treatments pertain only to the two extreme cases of relatively small (in the harmonic or quasiharmonic cases) and very large (the random coil calculation) conformational fluctuations, and are not applicable to states with intermediate chain flexibility.

More recently, powerful thermodynamic integration and perturbation techniques[1,19-22] have been introduced which enable one

[†] Biosym Technologies.

[‡] Supercomputer Computations Research Institute.

(1) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys., Biophys. Chem.* **1989**, *18*, 431.

(2) Roberts, V. A.; Dauber-Osguthorpe, P.; Osguthorpe, D. J.; Levin, E.; Hagler, A. T. *Isr. J. Chem.* **1986**, *27*, 198.

(3) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. *Proteins: Struct. Funct. Genet.* **1988**, *4*, 31.

(4) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

(5) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.

(6) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585.

(7) Gō, N.; Scheraga, H. A. *J. Chem. Phys.* **1969**, *51*, 4751.

(8) Gō, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535.

(9) Gō, N.; Lewis, P. N.; Scheraga, H. A. *Macromolecules* **1970**, *3*, 628.

(10) Gō, M.; Scheraga, H. A. *Biopolymers* **1984**, *23*, 1961.

(11) Flory, P. J. In *Principles of Polymer Chemistry*; Cornell University: New York, 1953; Chapter 14.

(12) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Wiley-Interscience: New York, 1969; p 32.

(13) Hagler, A. T.; Stern, P. S.; Sharon, R.; Becker, J. M.; Naider, F. *J. Am. Chem. Soc.* **1979**, *101*, 6842.

(14) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325.

(15) Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D. *Macromolecules* **1984**, *17*, 1370.

(16) Edholm, O.; Berendsen, H. J. C. *Mol. Phys.* **1984**, *51*, 1011.

(17) DiNola, A.; Berendsen, H. J. C.; Edholm, O. *Macromolecules* **1984**, *17*, 2044.

(18) Rojas, O. L.; Levy, R. M.; Szabo, A. *J. Chem. Phys.* **1986**, *85*, 1037.

(19) Kirkwood, J. G. In *Theory of Liquid*; Alder, B. J., Ed.; Gordon & Breach: New York, 1968.

0002-7863/92/1514-5386$03.00/0

to obtain not only a difference in free energy, $\Delta F$, of a conformational change, but also the difference in free energy due to chemical changes. Thus the relative stabilities of large systems such as protein–ligand complexes in water, and various mutations of amino acid residues, can be determined.[1,23-29] This category also includes the absolute free energy technique proposed by Stoessel and Nowak[30] (see also a related method by Rojas et al. in ref 18). In principle, these methods are rigorous. In practice, however, they are limited to relatively small conformational or chemical changes; otherwise, the required number of integration steps (known as "windows") becomes large and the process intractable. In addition, it is difficult to estimate the effect of the sample size at each integration step on the statistical error, which leaves uncertainty as to the precision of the result.[1]

An alternative approximate procedure for calculating the entropy, called the local states (LS) method, has been proposed by Meirovitch.[31] With this method the value of the sampling probability $P$ can be obtained approximately from the frequencies of occurrence of the so-called local states (for a polypeptide a local state is a set of values of neighboring dihedral and valence angles which define a local conformation of the chain). This procedure was originally developed to calculate the entropy of Ising and lattice gas systems simulated by the Monte Carlo technique;[31-33] later, it was extended to polymer chains[35-37] and to polypeptides.[34] In the latter study, the LS method was applied to a model of decaglycine without solvent, where the molecule, described by the potential energy function ECEPP,[38,39] has rigid geometry. It was found that the $\alpha$-helical state is more stable, i.e., has lower free energy, than the hairpin state.[34] The method has also been used recently in studies of the properties of elastin.[40] The LS method, in contrast to the methods mentioned above,[7,8,11,14] is general in the sense that it can be applied efficiently not only to stable states but also to the random coil and other chain flexibilities. This enables one to compare the stability of structures having large conformational differences.

In this work the LS method is further developed by applying it to the cyclic peptide cyclo-(Ala-Pro-D-Phe)$_2$ in order to assess the effect of environment on the free energy. Here, we also generalize the method to handle side chains in the local states definition, as well as extending it to calculate free energy from a system simulated with molecular dynamics; i.e., the molecule is described by flexible geometry. The structures of this molecule in the crystal and in solution have been determined by X-ray crystallography[41] and NMR.[42] This peptide has also been studied by theoretical procedures in vacuum (i.e., in the absence of solvent) and in the crystal by Hagler et al.[43,44] and Kitson and Hagler.[45,46] In the latter works one of the main objectives was to investigate the effects of environment on the properties of the peptide. This is important since biological processes, such as the binding of peptides to receptors, involve a change of environment. Here, we explore environmental effects further by estimating, for the first time, the reduction in the conformational entropy of cyclo-(Ala-Pro-D-Phe)$_2$ in going from vacuum to the crystal. For comparison we have also calculated this change in entropy using the harmonic approximation based on a normal mode analysis.[13]

## II. Theory

For the sake of simplicity the theory of the local states method will be described as applied to self-avoiding walks (SAWs) of $N$ steps (bonds) on a square lattice; thus, the excluded volume interaction is taken into account. The SAWs start from the origin of the lattice, and finite interactions (e.g., attractions) may be defined between the chain steps. Such lattice models have been used to describe proteins.[47-50] Only small modifications, which will be discussed in section II.E, are required for a continuum model of a peptide, i.e., a model in which the lattice restriction is removed.

**A. Thermodynamic Functions.** Let us define thermodynamic functions for self-avoiding walks on the lattice. The partition function, $Z$, is given by

$$Z = \sum_{i\in\Omega} \exp(-E_i/k_B T) \tag{1}$$

where $i$ runs over the ensemble $\Omega$ of all the possible configurations of the chain, $E_i$ is the energy of configuration $i$, $k_B$ is the Boltzmann constant, and $T$ is the absolute temperature. The Boltzmann probability of SAW $i$ is, therefore,

$$P_i^B = \exp(-E_i/k_B T)/Z \tag{2}$$

and the statistical average of any microscopic property $X_i$ (such as the energy $E_i$) is given by,

$$X = \langle X \rangle = \sum_{i\in\Omega} P_i^B X_i \tag{3}$$

The entropy, $S$, and the Helmholtz free energy, $F$, can be formally expressed as statistical averages,

$$S = -k_B \sum_{i\in\Omega} P_i^B \ln P_i^B \tag{4}$$

and, from eqs 3 and 4,

$$F = E - TS = \sum_{i\in\Omega} P_i^B E_i - T[-k_B \sum_{i\in\Omega} P_i^B \ln P_i^B] = \sum_{i\in\Omega} P_i^B [E_i + k_B T \ln P_i^B] \tag{5}$$

For long SAWs or realistic models of a polypeptide, *exact* calculation of the statistical average, $X$, of a property is intractable because of the tremendous number of possible configurations. However, $X$ can be estimated in the following way: one selects $n$ SAWs from the ensemble, according to their Boltzmann probability $P_i^B$ (eq 2), and then calculates the arithmetic average $\bar{X}$,

$$\bar{X} = n^{-1} \sum_{t=1}^{n} X_{i(t)} \tag{6}$$

where $i(t)$ is the $t$th SAW selected; for a large sample size $n$, $\bar{X}$

[20] Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.
[21] Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.
[22] Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245.
[23] Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284.
[24] Tembe, B. L.; McCammon, J. A. *Comput. Chem.* **1982**, *8*, 281.
[25] Straatsma, T. P.; Berendsen, H. J. C.; Postma, J. P. M. *J. Chem. Phys.* **1986**, *85*, 6720.
[26] Mezei, M. *J. Chem. Phys.* **1988**, *86*, 7084.
[27] Cieplak, P.; Kollman, P. A. *J. Am. Chem. Soc.* **1988**, *110*, 3734.
[28] Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050.
[29] Fleischeman, S. H.; Brooks, C. L., III *Proteins* **1990**, *7*, 52.
[30] Stoessel, J. P.; Nowak, P. *Macromolecules* **1990**, *23*, 1961.
[31] Meirovitch, H. *Chem. Phys. Lett.* **1977**, *45*, 389.
[32] Meirovitch, H. *J. Stat. Phys.* **1983**, *30*, 681.
[33] Meirovitch, H. *Phys. Rev. B* **1984**, *30*, 2866.
[34] Meirovitch, H.; Vāsquez, M.; Scheraga, H. A. *Biopolymers* **1987**, *26*, 651.
[35] Meirovitch, H. *Macromolecules* **1985**, *18*, 569.
[36] Meirovitch, H. *Phys. Rev. A* **1985**, *32*, 3709.
[37] Meirovitch, H.; Scheraga, H. A. *J. Chem. Phys.* **1986**, *84*, 6369.
[38] Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361.
[39] Sippl, M. J.; Nēmethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1984**, *88*, 6231.
[40] Wasserman, Z. R.; Salemme, F. R. *Biopolymers* **1990**, *29*, 1613.
[41] Brown, J. N.; Teller, R. G. *J. Am. Chem. Soc.* **1976**, *98*, 7565.
[42] Kopple, K. D.; Schamper, T. J.; Go, A. *J. Am. Chem. Soc.* **1974**, *96*, 2597.

[43] Hagler, A. T.; Moult, J. *Nature* **1978**, *272*, 222.
[44] Hagler, A. T.; Moult, J.; Osguthorpe, D. J. *Biopolymers* **1980**, *19*, 395.
[45] Kitson, D. H.; Hagler, A. T. *Biochemistry* **1988**, *27*, 5246.
[46] Kitson, D. H.; Hagler, A. T. *Biochemistry* **1988**, *27*, 7176.
[47] Taketomi, H.; Ueda, Y.; Gō, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
[48] Kolinski, A.; Skolnick, J.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 7267.
[49] Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986.
[50] Covell, D. G.; Jernigan, R. A. *Biochemistry* **1990**, *29*, 3287.

is expected to approach $X$. It should be pointed out that the Boltzmann probability, $P_i^B$, which appears in eq 3, does not appear in eq 6 because the self-avoiding walks, $i(t)$, in the sample are already distributed according to $P_i^B$. In order to understand the LS method, it is important to note the distinction between $X$ in eq 3 (the ensemble average) and its estimator $\bar{X}$ (eq 6); we shall always use the bar to denote an estimation. The accuracy of such an estimation depends on the standard deviation $\sigma_X$ of the property $X$

$$\sigma_X = [\sum_{i\in\Omega} P_i^B(X - X_i)^2]^{1/2} \qquad (7)$$

For extensive properties, such as the energy, $\sigma_X \sim N^{1/2}$, where $N$ is the number of steps in each SAW.[51] The convergence of $\bar{X}$ to $X$, however, is determined by $\sigma_{\bar{X}}$ which depends on both $N$, and the number of configurations in the sample, $n$;[52] for an *uncorrelated* sample (see below) it is $\sigma_{\bar{X}} = \sigma_x/n^{1/2} \approx (N/n)^{1/2}$, i.e., $\sigma_{\bar{X}}$ increases with the system size $N$ and decreases with the sample size $n$. The free energy $F$, in contrast to the energy, however, is a statistical average with zero fluctuations,[53] i.e., $\sigma_F = 0$.

Assume now two states of a molecule defined over different regions of phase space $\Omega_1$ and $\Omega_2$ (e.g., a helical and a $\beta$-turn state of a peptide) with partition functions $Z_1$ and $Z_2$, free energies $F_1$ and $F_2$, and average energies $E_1$ and $E_2$, respectively. In order to be able to estimate the difference in energy $\Delta E = E_1 - E_2$ with a good precision, one should generate two *uncorrelated* samples, each of size $n$ such that $\sigma_{\bar{E}_1}, \sigma_{\bar{E}_2} \ll \Delta E$ or $A(N/n)^{1/2} \ll \Delta E$, i.e., $n \gg A^2N/(\Delta E)^2$, where $A$ is an energy constant. For example, a molecular dynamics study of the protein *Streptomyces griseus* Protease A (SGPA)[54] (181 amino acid residues) shows that $\sigma_E$ (eq 7) is approximately 200 kcal/mol; this means that in order to obtain $\bar{E}$ with precision of ± 1 kcal/mol (i.e., $\sigma_{\bar{E}} = \sigma_E n^{-1/2} \sim 200n^{-1/2} < 1$ kcal/mol), an uncorrelated sample of $\sim$40 000 conformations is required. It should be noted, however, that the conformations in Monte Carlo and molecular dynamics samples are highly correlated, and, therefore, the sample size, $n$, required is much larger than the above estimate. (The correlation range for a quantity $X_i$ is defined as the time $t$ for which the autocorrelation function of $X_i$ becomes 0. Therefore, the number of uncorrelated conformations in a correlated sample may, to a good approximation, be given by $n' = n/t$ rather than $n$.) Indeed, for large protein–ligand complexes it was found impossible, because of the large energy fluctuations, $\sigma_E$, and because the samples were too small, to distinguish between the energies of different states (e.g., a mutated versus a wild type protein; see, for example, results for the energy in Table III of ref 29). On the other hand, *in principle*, the free energies, $F_1$ and $F_2$, of such states and hence the difference, $\Delta F = F_1 - F_2$, can be calculated *exactly* from two samples, each of size $n = 1$ only. As discussed in the Introduction, it is, however, difficult in most cases to calculate $F$ exactly. Thus, the local states method enables one to define an approximate free energy functional, $F^A$ (see below). The standard deviation $\sigma_F{}^A$ of this property will be larger than zero; however, $\sigma_F{}^A$, is expected to decrease monotonically as the approximation is improved; therefore $\sigma_F{}^A$ constitutes a criterion for the extent of approximation in $F^A$. Obviously, for a good enough approximation, one would expect $\sigma_F{}^A < \sigma_E$.

The local states method is based on the concepts of the scanning simulation method,[55,56] proposed by Meirovitch, which is a step-by-step procedure for generating polymer chains. We shall, therefore, first describe the scanning method.

**B. The Exact Scanning Procedure.** Unlike molecular dynamics[5,6] and Metropolis Monte Carlo,[4] which are dynamical type methods, i.e., the conformation of the system evolves with time in a correlated manner, the scanning method is a step-by-step construction procedure based on transition probabilities[55,56] in which the conformations in the sample are statistically independent. Let us describe the method as applied to self-avoiding walks of $N$ steps (without finite interaction energy) on a square lattice. The first step, which starts from the origin, is determined (using a random number) in one of the four possible directions, $\nu$, each with an equal probability of 1/4. In the next steps of the process ($k > 1$), the transition probability for selecting a direction $\nu_k$ becomes a function of the $k - 1$ previous steps, $\nu_1$, ..., $\nu_{k-1}$. We denote by $p(\nu_k|\nu_{k-1},...,\nu_1)$ the probability of selecting a direction $\nu_k$ at the $k$th step, given that the previous $k - 1$ steps are $\nu_1$, ..., $\nu_{k-1}$. These transition probabilities are calculated by generating (or "scanning") all of the possible so-called "future SAWs" that begin with each of the directions $\nu_k$. These future SAWs are the possible continuations of the chain to the full length $N$, in future steps (i.e., $k$, $k + 1$, ..., $N$). The number of future SAWs starting in each direction is counted and the larger this number for direction $\nu_k$ ($\nu_k$ = 1, 4), the larger is the value of $p(\nu_k|\nu_{k-1},...,\nu_1)$ (this procedure has been extended to models with interaction energy as well.) The direction $\nu_k$ is then selected with the help of a random number according to the transition probabilities and the process continues.

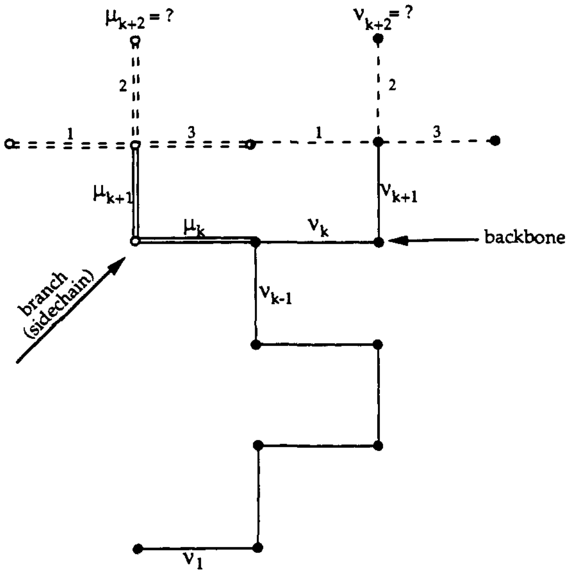

**Figure 1.** The possible directions available at step $k + 2$ from which a backbone direction $\nu_{k+2}$ and a branch direction $\mu_{k+2}$ will be selected simultaneously with the scanning procedure (the future self-avoiding-walks, $\nu_{k+3}$, ..., $\nu_N$ are not shown). The backbone and the branch are denoted by single and double lines, respectively. The branch is of $Q$ = 3 steps. The three possible directions of $\mu_{k+2}$ and $\nu_{k+2}$ are denoted by dashed lines. Obviously, the choice $\mu_{k+2}$ = 3 and $\nu_{k+2}$ = 1 is forbidden.

Once self-avoiding walk $i$ of $N$ steps has been constructed, one knows its construction probability $P_i$

$$P_i = P_i^B = \frac{1}{4}\prod_{k=2}^{N} p(\nu_k|\nu_{k-1},...,\nu_1) \qquad (8)$$

which is the product of the $N$ sequential transition probabilities with which the directions $\nu_1$, ..., $\nu_N$ have been chosen. One can show that $P_i$ (eq 8) is exact; i.e., it is equal to the Boltzmann probability[55,56] (eq 2). Since the number of SAWs increases exponentially with increasing $N$, the exact scanning procedure is impractical for large $N$.

The scanning method has also been extended to branched polymers.[57] If the branching points and the branch lengths are predefined (such as in a protein), the future scanning should encompass all possible backbone *and* branch conformations. In this case, however, the transition probabilities are more complex. If a branching point is reached at step $k$, the directions of two

[51] Hill, T. L. In *An Introduction to Statistical Thermodynamics*; Dover: New York, 1986.

[52] Binder, K. In *Monte Carlo Methods in Statistical Mechanics*; Binder, K., Ed.; Springer Verlag: Berlin, 1986; Chapter 1.

[53] Meirovitch, H.; Alexandrowicz, Z. *J. Stat. Phys.* **1976**, *15*, 123.

[54] Avbelj, F.; Moult, J.; Kitson, D. H.; James, M. N. G.; Hagler, A. T. *Biochemistry* **1990**, *29*, 8658.

[55] Meirovitch, H. *J. Phys. A* **1982**, *15*, L735.

[56] Meirovitch, H. *J. Chem. Phys.* **1988**, *89*, 2514.

[57] Meirovitch, H. *J. Phys. A* **1987**, *20*, 6059.

bonds, one of the backbone, $\nu_{k+q}$, and the other of the branch, $\mu_{k+q}$, should be determined at step $k + q$ $(0 \leq q \leq Q - 1)$ until the whole branch of $Q$ steps has been constructed (see Figure 1). However, if a new branching point happens to occur before the previous one has been completed, the backbone and the two branches will be generated simultaneously. For the case of a single branch, the transition probability at step $k + q$ is denoted by

$$p(\nu_{k+q},\mu_{k+q}|\nu_{k+q-1},...,\nu_1,\mu_{k+q-1},...,\mu_k) \qquad (9)$$

In summary, for our purposes it is important to realize that the scanning method is based on transition probabilities which depend on all the previous steps. In contrast to the Metropolis technique, where the value of $P_i^{\mathrm{B}}$ is unknown, the exact scanning method provides this value (eq 8), and hence the entropy, $S$, which is related to ln $P_i^{\mathrm{B}}$, is known as well (see eq 4). In this work we shall not apply the scanning method, but shall use its concepts as the basis for the local states method.

**C. The Local States Method: Application to Monte Carlo or Molecular Dynamics Realizations.** Suppose that the entropy is to be calculated from a sample of $n$ self-avoiding walks of $N$ steps each. Further assume that these $n$ configurations have been generated with the Metropolis method[4,58,59] (or, in principle, with any other simulation technique, such as molecular dynamics). The local states (LS) method is based on the fact that the properties of a *large* sample at equilibrium are independent of the simulation technique with which it has been generated. That is, although two equilibrium samples generated with different simulation techniques are very unlikely to be identical, they lead to estimates of average properties, $\bar{X}$ (see eq 6), that are equal to within a statistical error that decreases with increasing sample size $n$. Therefore one can *assume*[31-34] that the sample of SAWs has been generated with the scanning method, rather than by the actual Monte Carlo technique. This is the key to the local states method described here. This then enables one to reconstruct the transition probabilities of the scanning method from the frequencies of occurrence of the various configurations of the partial SAWs $(\nu_k, ..., \nu_1)$ $(k = 1, N)$. Such configurations are called local states.[31] Thus, for a given local state, one can count the number of times, $n(\nu_k, \nu_{k-1}, ..., \nu_1)$, that $\nu_k$ is preceded in the sample by the sequence $\nu_1...\nu_{k-1}$, and divide by $n(\nu_{k-1}, ..., \nu_1)$, the total number of occurrences of $\nu_1...\nu_{k-1}$, defining the transition probability,

$$p(\nu_k|\nu_{k-1},...,\nu_1) \cong n(\nu_k,\nu_{k-1},...,\nu_1)/n(\nu_{k-1},...,\nu_1) \qquad (10)$$

For a large sample [which means large values of $n(\nu_k, ..., \nu_1)$], the values of $p(\nu_k|\nu_{k-1}, ..., \nu_1)$ (eq 10) will approach the values of the corresponding transition probabilities of the exact scanning method. Having calculated the transition probabilities from the sample, the probability of each conformation can then be calculated using eq 8. However, reconstructing the transition probabilities of the exact scanning method is impractical for long chains because of the very large number of possible local states. Therefore, one can define approximate transition probabilities for the local states $(\nu_k, ..., \nu_{k-b})$ $(k = 1, N)$, which depend on only $b$ previous steps (rather than all $k - 1$), where $b$ is small ($b$ is called the correlation parameter), i.e.,

$$p(\nu_k|\nu_{k-1},...,\nu_{k-b}) = n(\nu_k,...,\nu_{k-b})/n(\nu_{k-1},...,\nu_{k-b}) \qquad (11)$$

Notice that for $k \leq b$ the transition probability is exact since all of the $(k - 1)$ previous steps are taken into account. [We shall also use approximations based on $b = 0$ (i.e., no correlations are taken into account); the corresponding transition probabilities are denoted by $p(\nu_k, b = 0)$.] Thus, the calculation of construction probabilities $P_i(b)$ (see eq 12 below) is carried out in two stages; first, a set of transition probabilities for the various local states is obtained and then the $N$ transition probabilities corresponding to configuration $i$ in the sample are determined and the approximate construction probability $P_i(b)$ assigned to it (see eq 8) is calculated

$$P_i(b) = \frac{1}{4}\prod_{k=2}^{N} p(\nu_k|\nu_{k-1},...,\nu_{k-b}) \qquad (12)$$

where the factor 1/4 is the transition probability for the first step on a square lattice with coordination number 4.

One also has to define approximate transition probabilities for branched chains. As has been discussed in section II.B, at step $k + q$, the directions $\nu_{k+q}$ and $\mu_{k+q}$ of the backbone and the branch, respectively, should be determined simultaneously. However, it is more convenient to define separate transition probabilities for the backbone and the branch. Thus, for the pairs $(\nu_k, \mu_k)$ and $(\nu_{k+1}, \mu_{k+1})$ the transition probabilities (for a given value of the correlation parameter $b$) can be factored as follows (using the Bayes formula):

$$p(\nu_k,\mu_k|\nu_{k-1},...,\nu_{k-b}) = p(\nu_k|\nu_{k-1},...,\nu_{k-b})p(\mu_k|\nu_k,...,\nu_{k-b}) \qquad (13)$$

$$p(\nu_{k+1},\mu_{k+1}|\nu_k,\mu_k,...,\nu_{k-(b+2)}) = p(\nu_{k+1}|\nu_k,\mu_k,...,\nu_{k-(b+2)})p(\mu_{k+1}|\nu_{k+1},\mu_k,\nu_k,...,\nu_{k-(b+2)}) \qquad (13a)$$

where a similar factorization can be defined for a general pair of backbone and branch steps, $(\nu_{k+q}, \mu_{k+q})$. For cyclo-(Ala-Pro-D-Phe)$_2$ treated here, the values of $b$ studied are relatively small, $b \leq 3$ for the backbone angles and $b \leq 1$ for the side chains. We also use an approximation in which correlations are not taken into account (i.e., $b = 0$, for both backbone and side chains; see section III.A).

**D. Approximations of the Free Energy.** Having described the way in which the approximate construction probabilities, $P_i(b)$, can be calculated, we shall now discuss approximations to the free energy that can be obtained from these probabilities. We should first note that with the exact scanning method the whole future is searched, and therefore self-intersecting chains cannot be generated. However, a step-by-step construction procedure based on the approximate transition probabilities (eq 11) can lead to self-intersecting random walks (i.e., the chain can fall on itself, which is an unphysical state). This means that the approximate probability $P_i(b)$ defined above (eq 12) is not normalized over the ensemble of SAWs $\Omega$, but over a larger ensemble that also includes self-intersecting configurations. Therefore, $G$, defined by

$$G = \sum_{i\in\Omega} P_i(b) < 1 \qquad (14)$$

is less than 1 and a normalized probability, $P_i(b)$, for the whole ensemble of self-avoiding walk configurations $\Omega$ is

$$P_i'(b) = P_i(b)/G \qquad (15)$$

If a step-by-step construction procedure based on the approximate transition probabilities given by eq 11 is used to generated SAWs, one can estimate $G$ by the ratio of the number of SAWs completed successfully to the number started. However, $G$ cannot be obtained from a sample generated by the Monte Carlo method since this sample includes only "successful" SAWs. We therefore study the entropy functional $S^{\mathrm{A}}$ which depends on $P_i(b)$ rather than on $P_i'(b)$,

$$S^{\mathrm{A}} = -k_{\mathrm{B}}\sum_{i\in\Omega} P_i^{\mathrm{B}} \ln P_i(b) \qquad (16)$$

Using Jensen's inequality,[36,60] $S^{\mathrm{A}}$ can be shown rigorously to overestimate the correct entropy $S$ (eq 4), i.e.,

$$S^{\mathrm{A}} \geq S \qquad (17)$$

The corresponding free energy $F^{\mathrm{A}}$ is based on the correct average energy $E$ (eq 3) and $S^{\mathrm{A}}$,

$$F^{\mathrm{A}} = \sum_{i\in\Omega} P_i^{\mathrm{B}}[E_i + k_{\mathrm{B}}T \ln P_i(b)] = E - TS^{\mathrm{A}} \qquad (18)$$

which, due to eqs 5 and 17, satisfies the relation

$$F^{\mathrm{A}} \leq F \qquad (19)$$

$S^{\mathrm{A}}$ (eq 16) and $F^{\mathrm{A}}$ (eq 18) are statistical averages defined with the Boltzmann probability $P_i^{\mathrm{B}}$; this definition is chosen because

[58] Verdier, P. H.; Stockmayer, W. H. *J. Chem. Phys.* **1962**, *36*, 227.
[59] Kremer, K.; Baumgärtner, A.; Binder, K. *J. Phys. A.* **1982**, *15*, 2879.
[60] Prazen, E. *Modern Probability Theory and its Application*; Wiley: New York, 1968.

the sample is generated with $P_i^B$ by the Monte Carlo procedure. Thus $F^A$ can be obtained from a sample of $n$ chains generated by the Monte Carlo procedure by calculating the arithmetic average $\bar{F}^A$,

$$\bar{F}^A = n^{-1}\sum_{t=1}^{n} E_{i(t)} + k_BT \ln P_{i(t)}(b) = \bar{E} - T\bar{S}^A \quad (20)$$

Notice that the transition from the ensemble average $F^A$ (eq 18) to its estimate $\bar{F}^A$ is the same as that from $X$ to $\bar{X}$ in eqs 3 and 6, respectively.

Another free energy functional, $F^B$, has also been introduced[35]

$$F^B = \sum_{i\in\Omega} P_i'(b)[E_i + k_BT \ln P_i(b)] = E^B - TS^B \quad (21)$$

This quantity is expected to provide an upper bound for the correct free energy $F$ as explained below. If $\ln P_i(b)$ in eq 21 is replaced by $\ln P_i'(b)$ (eq 15), the natural log of the normalized probability $P_i'(b)$, $F^B$ becomes a free energy functional, $F^B(P_i'(b))$, defined with a *single* probability distribution $P_i'(b)$ [rather than two different probability distributions used in defining both $F^A$ (eq 18) and $F^B$ (eq 21)]. It is known from statistical mechanics (the minimum free energy principle[61,62]) that such a functional, i.e., $\sum P_i'(b)$ $[E_i + k_BT \ln P_i'(b)]$ is minimal for the Boltzmann probability, i.e., when $P_i'(b) = P_i^B$. Therefore, $F^B(P_i'(b))$ overestimates the correct $F$ (eq 5), i.e., $F^B(P_i'(b)) \geq F$. However, as has already been pointed out, it is impossible to obtain the normalized probability $P_i'(b)$ from the Monte Carlo sample, and therefore $F^B(P_i'(b))$ cannot be calculated as well. On the other hand, $F^B$ (eq 21) defined with $\ln P_i(b)$ can be calculated; however, for this quantity it can only be shown[36] that

$$F^B \geq F^A \quad (22)$$

However, in practice $F^B$ still constitutes a very useful quantity since for all the models studied thus far $F^B$ has been found to be an upper bound of the correct value,[34-37] $F$. In part of these systems, i.e., SAWs on a square and a simple cubic lattice (which are also enclosed in a small volume[36]) and the continuum model of a freely jointed chain of hard disks,[37] the correct free energy $F$ has been obtained by other techniques. However, it should be pointed out that in order to determine that $F^B$ is an overestimation, one does not necessarily have to know the value of $F$. This can be achieved by verifying that the results for $F^B$ decrease monotonically as the approximation improves (i.e., as the correlation parameter $b$ is increased). In fact, the above-mentioned studies showed that, for good enough approximations, $F^B$ and $F^A$ deviate approximately equally from the correct free energy $F$ (eq 5). In this case their average, $F^M$,

$$F^M = (F^A + F^B)/2 \quad (23)$$

becomes a better approximation for $F$ than either one of them individually.

**1. Estimation of $F^B$ by Importance Sampling.** The estimation of the approximate free energy $F^B$ has been discussed before.[34,36,37] $F^B$, in contrast to $F^A$, is a statistical average defined with the probability $P_i'(b)$, which differs from the sampling probability $P_i^B$. Therefore, $F^B$ cannot be estimated straightforwardly from the sample in the same manner as $F^A$ (see eq 20). $F^B$ can be estimated,[34-37] however, in a more elaborate way by two procedures, importance sampling[63,64] and the generalized Monte Carlo procedure[4,52] suggested by Schmidt.[65] To apply importance sampling (IS), one has first to express $F^B$, which is defined with the approximate probability $P_i'(b)$, as an average defined with the Boltzmann probability $P_i^B$ (which is the sampling probability). This average can be estimated from the Monte Carlo sample of size $n$ by (see refs 34-37),

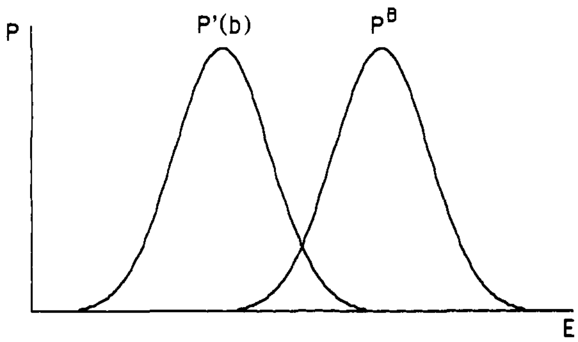

Figure 2. Schematic representation, as a function of energy $E$, of the exact Boltzmann probability, $P^B$ (eq 2), and the approximate probability $P'(b)$ (eq 15). The most probable conformations with respect to $P^B$ belong to the tail of $P'(b)$. (Note that the actual probability distributions will not conform to the symmetrical Gaussian distributions illustrated here.)

$$\bar{F}^B(\mathrm{IS}) = \sum_{t=1}^{n} P_{i(t)}(b)\exp(E_{i(t)}/k_BT)[E_{i(t)} + k_BT \ln P_{i(t)}(b)] \Big/ \sum_{t=1}^{n} P_{i(t)}(b)\exp(E_{i(t)}/k_BT) \quad (24)$$

where the transformation from $F^B$ to $\bar{F}^B$(IS) is like that of $X$ (eq 3) to $\bar{X}$ (eq 6). Thus, for each configuration $i(t)$ of the Monte Carlo sample one knows the energy $E_{i(t)}$ and the probability $P_{i(t)}(b)$. The expressions in the numerator and denominator are calculated and summed up to give $\bar{F}^B$(IS).

It should be emphasized that the convergence of $\bar{F}^B$(IS) to $F^B$ is, in general, much slower than that of $\bar{F}^A$ to $F^A$ (eqs 18 and 20), which means that a larger sample $n$ is required for $\bar{F}^B$(IS) than for $\bar{F}^A$. This stems from the fact that the expressions in the numerator and denominator of eq 24 are not extensive [i.e., they are not proportional to the number of steps (or atoms) $N$, such as for a usual thermodynamics quantity (e.g., the energy), since they consist of the probability $P_i(b)$ and the exponential term, which behaves as $\exp(-N)$]. If the distributions $P_i^B$ and $P_i'(b)$ differ significantly, very large sample size $n$ is needed for $\bar{F}^B$(IS) to converge to $F^B$ (see discussion following eq 7). This stems from the fact that the conformations which contribute significantly to $\bar{F}^B$(IS) (i.e., have large values of $P_{i(t)}'(b)$) (recall that the probability distribution $P_i'(b)$ is related to $P_i(b)$, which appears in eq 24, by the factor $G$ (eq 15)) belong to the tail of the sampling probability distribution, $P_i^B$. As $P_i(b) \rightarrow P_i^B$, the estimation of $F^B$ by $\bar{F}^B$(IS) becomes more efficient since more conformations in the sample contribute significantly to $F^B$ (see Figure 2). The effective sample size for $F^B$ can be obtained by the Schmidt procedure discussed in the next section.

**2. Estimation of $F^B$ with the Schmidt Procedure.** In the previous section we have shown that $F^B$ can be estimated by expressing it as a statistical average defined with the sample probability $P_i^B$ (eq 24) rather than $P_i'(b)$ (eq 21). With the Schmidt procedure, on the other hand, the process is reversed; i.e., one first changes the original Monte Carlo sample to one distributed with the approximate probability $P_i'(b)$ from which $F^B$ can be estimated directly using eqs 21 and 6. Thus, the Schmidt procedure[65] enables one to extract from a sample generated with the correct $P_i^B$ (the unbiased sample) an *effectively* small sample selected with $P_i'(b)$ (the biased sample). That is, a sample is obtained that is representative of that which would have been obtained if the original simulation had generated the sample according to $P_i'(b)$ rather than $P_i^B$. $F^B$ can then be estimated straightforwardly from the biased sample. The procedure is carried out as follows. The first conformation $i(t = 1)$ of the given unbiased Monte Carlo sample is always accepted for the biased sample. The second conformation $i(t = 2)$ is accepted with

[61] Gibbs, J. W. In *Elementary Principles in Statistical Mechanics*; Yale University Press: New Haven, CT, 1902; Chapter XI.

[62] Huber, A. In *Mathematical Methods in Solid State and Superfluid Theory*; Clark, R. C., Dermick, G. H., Eds.; Plenum Press: New York, 1968.

[63] Kahn, H. In *Symposium on Monte Carlo Methods*; Meyer, H. A., Ed.; Wiley: New York, 1956; p 146.

[64] Hammersley, J. M.; Handscomb, D. C. In *Monte Carlo Methods*; Methuen: London, 1964; p 57.

[65] Schmidt, K. E. *Phys. Rev. Lett.* **1983**, *51*, 2175.

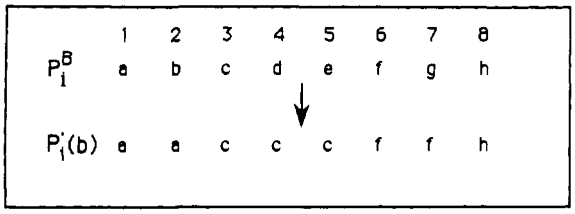

**Figure 3.** The Schmidt procedure. An original sample of $n = 8$ different conformations has been obtained with an exact probability $P_i^B$ (eq 2). After carrying out the Schmidt procedure, only $n_{\text{accept}} = 4$ different conformations, a, c, f, and h, have been accepted and they appear with some degeneracy. For large $n$, the accepted sample would be selected with the approximate probability $P_i'(b)$ (eq 15).

probability $A$. Let us define $A$ for general steps $t$ and $t + 1$. First, we define $\Delta E(t, t + 1)$ as the difference in energy between successive conformations in the sample divided by $k_BT$ (notice an error in ref 34, eqs 33 and 34)

$$\Delta E(t,t+1) = (E_{i(t+1)} - E_{i(t)})/k_BT \qquad (25)$$

Then

$$A[i(t),i(t+1)] = \min\{1,\exp[\Delta E(t,t+1)]P_{i(t+1)}(b)/P_{i(t)}(b)\} \qquad (26)$$

where $P_{i(t+1)}(b)$ and $P_{i(t)}(b)$ are the probabilities of the successive configurations obtained with eqs 11 and 12; $b$ is the correlation parameter. Equation 26 describes a generalized Metropolis Monte Carlo procedure which satisfies the detailed balance condition. This equation is expressed in terms of $P_i(b)$ rather than $P_i'(b)$ since the factor $G$ cancels out in the ratio; see eq 15. The decision as to whether to accept or reject $i(t = 2)$ is made with the help of a random number. If $i(t = 2)$ is accepted (i.e., $A$ is not smaller than the random number), the biased sample consists of the two conformations $i(t = 1)$ and $i(t = 2)$; if the conformation $i(t = 2)$ is rejected, the biased sample has the same conformation $i(t = 1)$ twice. This process continues for $t = 3$, etc., until all the conformations of the original unbiased sample have been tested and thus a biased sample [generated with $P_i'(b)$] is extracted from the unbiased one (see Figure 3). This procedure differs from a usual Metropolis procedure[4,52] in two respects. First, in a usual procedure one seeks to simulate the system with the Boltzmann probability $P_i^B$, whereas here we want the sample to be distributed according to the biased probability $P_i'(b)$; this gives rise to the following ratio $P_{i(t+1)}(b)/P_{i(t)}(b)$ in eq 26 instead of $P^B_{i(t+1)}/P_{i(t)}$ (=exp[$-\Delta E(t,t+1)$]), which appears in a usual Metropolis procedure. Second, the expression $\exp[\Delta E(t,t+1)]$ [$(=P^B_{i(t)}/P^B_{i(t+1)})$] (i.e., the ratio of the Boltzmann probabilities, eq 2) that appears in eq 26 is replaced by 1 in a usual Metropolis procedure. It is required here in order that the procedure satisfies the detailed balance condition, which guarantees the correct convergence. The efficiency of producing accepted conformations can be measured by the acceptance rate $R_a$,

$$R_a = n_{\text{accept}}/n \qquad (27)$$

where $n_{\text{accept}}$ is the number of chains accepted in the process. $R_a$ would be expected to increase as the approximation improves; for the exact Boltzmann probability $P_i^B$, $R_a = 1$, i.e., the whole original sample is accepted. This can be seen by expressing $P_{i(t+1)}(b)$ and $P_{i(t)}(b)$ in eq 26 in terms of the energies $E_{i(t+1)}$ and $E_{i(t)}$ (see eq 2), which gives rise to $A = 1$ for all $t$. Since $n_{\text{accept}} < n$, the size of the original Monte Carlo sample, some SAWs in the accepted sample will appear more than once (on average an accepted SAW will be repeated $n/n_{\text{accept}}$ times). In the initial stage of the process, the acceptance rate $R_a$ is relatively large since the first SAW considered, which is always accepted, is likely to be highly probable with $P_i^B$. It therefore will belong to the tail of the approximate probability $P_i'(b)$ (see discussion on $F^B$ in section III.A and Figure 2). As the process continues, structures that are more probable with respect to $P_i'(b)$ are accepted, and therefore the acceptance rate $R_a$ decreases until it becomes stable. This means that the SAWs accepted prior to the stabilization should be discarded. The free energy functional $F^B$ is a statistical average defined with the biased probability $P_i'(b)$ (eqs 15 and 21); it can, therefore, be estimated from the biased sample of accepted chains (selected with $P_i'(b)$) by the estimator $\bar{F}^B$(A) [where (A) stands for "accepted"] in the same manner as the free energy functional $F^A$ is estimated by $\bar{F}^A$ (see eqs 18 and 20)

$$\bar{F}^B(\text{A}) = n^{-1}\sum_{t=1}^{n}{}'E_{i(t)} + k_BT \ln P_{i(t)}(b) \qquad (28)$$

where $\sum'$ denotes summation over the accepted biased sample. The effective sample size for $\bar{F}^B$(A), which determines the statistical error, is the number of accepted SAWs, $n_{\text{accept}}$. This should be compared to the larger sample size, $n$, available for $F^A$. Schmidt's procedure, as well as importance sampling, is efficient only if the probabilities $P_i'(b)$ and $P_i^B$ are sufficiently close to each other (see Figure 2). Otherwise, the acceptance rate becomes very small, which means that the sample size $n$ needs to be extremely large.

To summarize, the approximate free energies $F^A$ and $F^B$ are expected to approach $F$, the correct free energy, from opposite sides as the approximation improves, i.e., as $P_i'(b) \rightarrow P_i^B$, providing lower and upper bounds for $F$. In practice, one calculates several approximations by using different values of the correlation parameter, $b$, and studies the convergence of the results for $F^A$, $F^B$, and $F^M$, from which the best estimate of $F$ is obtained. The value of $n_{\text{accept}}$ obtained from the Schmidt procedure constitutes a convenient measure for the statistical reliability of the estimate of $F^B$.

Another measure of the extent of convergence of $F^A$ is its fluctuation (eq 7) $\sigma_F{}^A$, i.e.,

$$\sigma_F{}^A = \{\sum_{i\in\Omega} P_i{}^B[F^A - E_i - k_BT \ln P_i(b)]^2\}^{1/2} \qquad (29)$$

where $F^A$ is the average value (eq 18), and the expression in square brackets is the deviation from the average for each $i$. We have already pointed out that the fluctuation of the exact free energy $F$ is zero, whereas, generally, $\sigma_F{}^A > 0$. Therefore, one would expect that $\sigma_F{}^A \rightarrow 0$ as the approximation improves. $\sigma_F{}^A$ can be estimated by $\bar{\sigma}_F{}^A$, from the original Monte Carlo sample of size $n$ where $F^A$ in eq 29 becomes $\bar{F}^A$ (eq 20).

$$\bar{\sigma}_F{}^A = \left\{\frac{1}{n}\sum_{t=1}^{n}[\bar{F}_A - E_{i(t)} - k_BT \ln P_{i(t)}(b)]^2\right\}^{1/2} \qquad (30)$$

The various free energy and entropy functionals described above are defined for discrete models. However, they also apply to continuum models (discussed in the next section) where the only change is that the probabilities $P_i$ are replaced by probability densities $\rho$ and the summations over the ensemble of SAWs become integrations over a continuum phase space $\Omega$. However, notice that the estimators (eqs 20, 24, 28, and 30) are still defined with summations where $\rho$ replaces $P$.

**E. The Local States Method for Polypeptides.** Assume first a polypeptide model of $N$ amino acid residues having a rigid geometry, i.e., constant bond lengths and bond angles. A conformation is defined by the $K$ backbone and side chain dihedral angles $\phi$, $\psi$, $\omega$, and $\chi$, denoted here by $\alpha_k$, $1 \le k \le K$, which can have continuum values in the range [−180°, 180°]. In order to apply the LS method, one has to divide this region into a discrete number of segments in the following way. First, from a sample of polypeptide conformations one calculates the ranges $\Delta\alpha_k$, within which the values of $\alpha_k$, $1 \le k \le K$, lie (again $\alpha_k$ represents $\phi$, $\psi$, $\omega$, and $\chi$),

$$\Delta\alpha_k = \alpha_k(\text{max}) - \alpha_k(\text{min}) \qquad (31)$$

where $\alpha_k$(max) and $\alpha_k$(min) are the maximal and minimal values of $\alpha_k$ found in the sample. In the next stage $\Delta\alpha_k$ is divided into $l$ equal segments of sizes $\Delta\alpha_{k,l}$,

$$\Delta\alpha_{k,l} = \Delta\alpha_k/l \qquad (32)$$

We denote these segments of dihedral angles by $\nu_k$ ($\nu_k = 1, l$), $1 \le k \le K$ (notice that for SAWs on a square lattice, $l$ is constant; i.e., $l$ is equal to the coordination number 4). For a given conformation $i$ one can find the particular set of segments (which constitutes a $K$-dimensional vector) $(\nu_{1(i)}, ..., \nu_{K(i)})$ to which the $K$ dihedral angles $\alpha_k$ belong. In a similar way one can also define $b$-dimensional vectors based on subgroups of $b$ consecutive angles only, where $b$, the correlation parameter, is smaller than $K$. The latter vectors are called local states and they correspond to the local states defined for SAWs in section III.C; the related transition probabilities are defined in the same way as for SAWs. The only significant difference here is the fact that for a continuum space one has to define transition probability densities $\rho$ (see later in this section), i.e., to divide the transition probability by the segment size, $\Delta\alpha_{k,l}$ (eq 32)

$$\rho(\nu_k|\nu_{k-1},...,\nu_{k-b}) = n(\nu_k,...,\nu_{k-b})/[n(\nu_{k-1},...,\nu_{k-b})\Delta\alpha_{k,l}] \qquad (33)$$

(i.e., according to eq 11, $\rho$ is $p(\nu_k|\nu_{k-1},...,\nu_{k-b})/\Delta\alpha_{k,l}$). The approximate probability density for the whole conformation is

$$\rho([\alpha_k], b, l) = \prod_{k=1}^{K} \rho(\nu_k|\nu_{k-1},...,\nu_{k-b}) \qquad (34)$$

where $[\alpha_k]$ denotes the set of values of $\alpha_k$, $1 \leq k \leq K$; each torsional angle $\alpha_k$ lies within the segment $\nu_k$, respectively.

Suppose now that a polypeptide molecule is simulated by molecular dynamics; i.e., it is modeled with flexible geometry. In this case one can still define a conformation, to a very good approximation, by its backbone and side chain dihedral angles $\phi$, $\psi$, $\omega$, and $\chi$, and also the bond angles $\theta$ and the bond lengths. However, the contribution of bond stretching to the entropy is clearly negligible[66,14] as the frequencies of these modes are much higher than $k_BT/h$ where $h$ is the Planck constant. Therefore, this contribution will be neglected in this work; i.e., we assume that the bond lengths are constant. Thus, as for the case of rigid geometry, one can define a set of $\Delta\alpha_k$ (eq 31), which now also includes the valence angles $\theta$ (the total number of angles, $K$, now includes these valence angles). An order among the backbone and side chains angles is then defined, which leads to the definition of various local states and transition probability densities (see eqs 33 and 34). Notice that these probability densities take into account correlation between the $b$ successive angles defining a local state, where $b$ is the correlation parameter.

The above application of the LS method to continuum models using internal coordinates is based on theoretical aspects that ought to be discussed. For such models the conformational partition function $Z$ (eq 1) (which is a summation in the case of a discrete model) is an integral of the function $\exp(-E/k_BT)$ with respect to the Cartesian coordinates over the whole phase space. For a stable state (e.g., an $\alpha$-helical state) the integration, however, is carried out over the limited region $\Omega$ that defines the state. This partition function leads to the average energy which is estimated from the molecular dynamics sample.

In order to obtain the entropy, one has first to change the variables of integration in $Z$ from Cartesian to internal coordinates, which makes the integrand more complex, i.e., dependent also on the Jacobian. However, the Jacobian $J$ has been shown to be a simple function of the bond lengths and angles (but not of the dihedral angles), and, therefore, if the potentials of these "hard variables" are strong, the average values of the variables can be assigned to $J$, which to a good approximation, can be taken out of the integral (see refs 8, 14, and 67). One has also to assume that the bond lengths are not correlated with the bond and dihedral angles, which means, for example, that the effect of the "cross terms" in the Hamiltonian is ignored. This enables one to carry out the integration over the bond lengths and the remaining integral becomes a function of the dihedral and valence angles only; hence if the angles are expressed in radians, the integral is dimensionless. One obtains for the partition function $Z'$

$$Z' = DZ = D \int \exp\{-E([\alpha_k])/k_BT\} d\alpha_1 \ldots d\alpha_K \qquad (35)$$

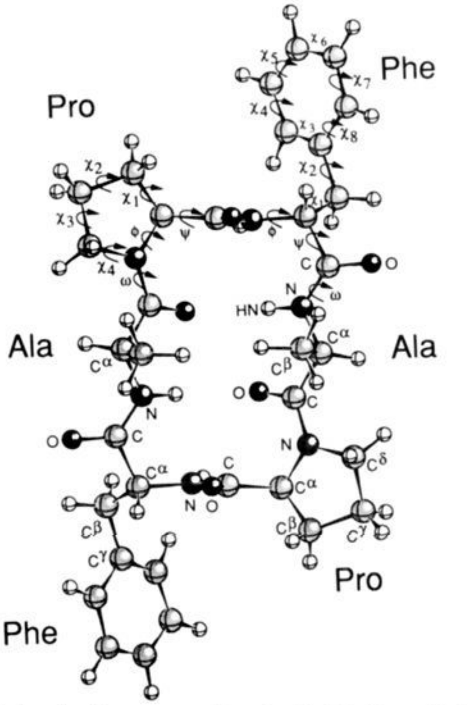

Figure 4. Ball-and-stick representation of cyclo-(Ala-Pro-D-Phe)$_2$. Hydrogen, carbon, nitrogen, and oxygen atoms are shown in open, light gray, medium gray, and dark gray balls, respectively.

The prefactor $D$ is a product of $J$ and the integral over the bond lengths; it depends on $T$ and the units in which the bond lengths are expressed. Obviously, in order to calculate the absolute free energy or entropy one needs to know $D$. However, if one is only interested in the differences $\Delta F$ or $\Delta S$ between two states of the same molecule at the same $T$, $D$ cancels out and, therefore, one can assume that $D = 1$. The corresponding Boltzmann probability density is (compare with eq 2)

$$\rho([\alpha_k]) = \exp[-E([\alpha_k])/k_BT]/Z \qquad (36)$$

and the entropy is

$$S = -k_B \int \rho([\alpha_k]) \ln \rho([\alpha_k]) d\alpha_1 \ldots d\alpha_K \qquad (37)$$

This approximate entropy is estimated with the LS method. It should be pointed out, however, that for convenience, in our calculations the angles are expressed in degrees rather than radians, and therefore $S$ (eq 37) is defined up to an additive constant.

## III. Results and Discussion

In this work we studied the polypeptide cyclo-(Ala-Pro-D-Phe)$_2$ (Figure 4) in vacuum and in the crystal. The two systems were simulated by molecular dynamics (at $T \sim 300$ K). The crystal simulation (of the whole unit cell, which consists of two peptide molecules and 16 water molecules) was started from the X-ray structure of the peptides (the waters were placed at random, but in sterically reasonable locations). All hydrogen atoms were included explicitly and a full valence force field, including cross terms, was used. The SPC water model,[68] with the addition of flexibility, was used to model water interactions. Periodic boundary conditions and a 15-Å cutoff distance were used for the crystal simulation. The force field and parameters are described in detail by Kitson and Hagler.[45] Every 8 femtoseconds (fs) the Cartesian coordinates of the system were stored in a file for later use. In this way two samples were created, one for the isolated molecule in vacuum (of size $n = 112\,500$ conformations, obtained from a 900-ps trajectory) and the other for the unit cell ($n = 36\,250$, obtained from a 290-ps trajectory). However, because

[66] Dauber, P.; Osguthorpe, D. J.; Sharon, R.; Stern, P.; Goodman, M.; Hagler, A. T. Computer Simulation of Biomolecular Systems. In *Proceedings of ACS Symposium on Supercomputers in Chemistry*, 1981, p 161.

[67] Flory, P. J. *Macromolecules* **1974**, *7*, 381.

[68] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, Holland, 1981; p 331.

**Table I.** The Ranges[a] within Which the Dihedral and Valence Angles Change during the Molecular Dynamics Simulations of Cyclo-(Ala-Pro-D-Phe)$_2$

Vacuum[b,d]

| residue no. | residue | $\Delta\phi$ | $\Delta\psi$ | $\Delta\omega$ | $\Delta\theta$(N–C$^\alpha$–C′) | $\Delta\theta$(C$^\alpha$–C′–N) | $\Delta\theta$(C′–N–C$^\alpha$) |
|---|---|---|---|---|---|---|---|
| 1 | Ala | 152 | 131 | 90 | 33 | 24 | 21 |
| 2 | Pro | 114 | 137 | 95 | 32 | 25 | 22 |
| 3 | D-Phe | 144 | 193 | 75 | 30 | 25 | 23 |
| 4 | Ala | 153 | 128 | 95 | 30 | 23 | 20 |
| 5 | Pro | 113 | 125 | 94 | 34 | 26 | 23 |
| 6 | D-Phe | 146 | 197 | 76 | 33 | 26 | 23 |

| | | $\chi^1$ | $\chi^2$ | $\chi^3$ | $\chi^4$ | $\chi^5$ | $\chi^6$ | $\chi^7$ | $\chi^8$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Ala | 230 | | | | | | | |
| 2 | Pro | 102 | 109 | 107 | 108 | | | | |
| 3 | D-Phe | 173 | 360 | 60 | 53 | 50 | 49 | 49 | 53 |
| 4 | Ala | 360 | | | | | | | |
| 5 | Pro | 103 | 113 | 106 | 93 | | | | |
| 6 | D-Phe | 173 | 360 | 60 | 50 | 50 | 50 | 50 | 60 |

Crystal[c,d]

| residue no. | residue | $\Delta\phi$ | $\Delta\psi$ | $\Delta\omega$ | $\Delta\theta$(N–C$^\alpha$–C′) | $\Delta\theta$(C$^\alpha$–C′–N) | $\Delta\theta$(C′–N–C$^\alpha$) |
|---|---|---|---|---|---|---|---|
| 1 | Ala | 97 | 60 | 62 | 23 | 22 | 20 |
| 2 | Pro | 76 | 79 | 49 | 31 | 25 | 19 |
| 3 | D-Phe | 92 | 94 | 61 | 24 | 24 | 22 |
| 4 | Ala | 116 | 61 | 60 | 26 | 20 | 20 |
| 5 | Pro | 87 | 81 | 50 | 29 | 26 | 21 |
| 6 | D-Phe | 98 | 87 | 62 | 25 | 24 | 21 |

| | | $\chi^1$ | $\chi^2$ | $\chi^3$ | $\chi^4$ | $\chi^5$ | $\chi^6$ | $\chi^7$ | $\chi^8$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Ala | 360 | | | | | | | |
| 2 | Pro | 53 | 72 | 76 | 70 | | | | |
| 3 | D-Phe | 67 | 70 | 53 | 48 | 48 | 51 | 47 | 52 |
| 4 | Ala | 360 | | | | | | | |
| 5 | Pro | 84 | 102 | 97 | 79 | | | | |
| 6 | D-Phe | 50 | 75 | 57 | 49 | 47 | 54 | 47 | 51 |

[a] The ranges are defined in eq 31, and their values in degrees are rounded off to the nearest whole number. [b] Results for the polypeptide in vacuum. [c] Results for the polypeptide chains in the crystal based on the maximum and minimum values of the angles for the two peptides in the unit cell. [d] The results for the valence angles of the side chains are 15° $\leq \Delta\theta \leq$ 30° (15° $\leq \Delta\theta \leq$ 20° for the D-Phe rings) for both the vacuum and the crystal. The corresponding values in the two environments do not differ by more than 5°.

only 8 fs separates consecutive conformations, they are probably correlated. These trajectories have been obtained by continuation of those studied by Kitson and Hagler.[45,46] In the first stage of application of the local states method, each peptide conformation in the sample was expressed in terms of a set of 92 backbone and side chain dihedral angles $\phi$, $\psi$, $\omega$, and $\chi$ and the valence angles $\theta$. From these sets of angles the entropy and free energy were calculated.

**A. The Peptide in Vacuum.** We shall now describe in detail the local states method as applied to the peptide in vacuum. First, let us specify the dihedral and valence angles defined for each conformation. There are the 3 × 6 = 18 backbone dihedral angles $\phi$, $\psi$, and $\omega$ for the six amino acid residues of the molecule. The number of angles $\chi$ of the side chains is one for Ala, four for Pro, and eight for D-Phe (we define dihedral angles also for the phenyl ring because of the relatively large changes observed for these angles, see below). A total of 48 valence angles $\theta$ are defined for the backbone atoms, N, C$^\alpha$, and C′ and for the side chain carbons. Valence angles for the hydrogens or oxygens are not taken into account. Therefore, the total number of angles treated is $K$ = 92.

To apply the LS method one first has to define a hypothetical approximate scanning buildup procedure for constructing the polypeptide. We assume that backbone construction starts from one of the alanine residues denoted by Ala$^1$; i.e., for each conformation the N–C$^\alpha$ bond of Ala$^1$ is considered to be fixed in space, thereby constituting a "seed" for the step-by-step construction (see Figure 4). Since the backbone forms a closed ring, the values of $\omega$ and $\theta$(C′–N–C$^\alpha$) of Phe$^6$, the last two dihedral angles in the ring, are fully determined by the values of the other backbone dihedral angles (see Figure 4); therefore, with such a procedure, the transition probabilities of these two angles are considered to be 1, which means that their contribution to the entropy is zero. Also, after the backbone valence angles $\theta$(N–C$^\alpha$–C′) (and $\theta$(C′–N–C$^\alpha$) of proline) have been determined, the directions of the C$^\alpha$–C$^\beta$ bonds (and also N–C$^\delta$ of proline) are almost completely defined. Therefore, the transition probabilities for the side chain valence angles $\theta$(N–C$^\alpha$–C$^\beta$) for all residues and those for $\theta$(C$^\delta$–N–C$^\alpha$) of proline are also ignored. For the proline side chain we define a scanning construction which goes from C$^\alpha$ to C$^\delta$. Again, because of the topology of the proline ring (see discussion for the backbone ring above), the values of $\chi^3$, $\chi^4$ and $\theta$(N–C$^\alpha$–C$^\beta$), $\theta$(C$^\gamma$–C$^\delta$–N), and $\theta$(C$^\delta$–N–C$^\alpha$) are already determined at this stage, and therefore only the four variables $\chi^1$, $\chi^2$, $\theta$(C$^\alpha$–C$^\beta$–C$^\gamma$), and $\theta$(C$^\beta$–C$^\gamma$–C$^\delta$) are considered (see Figure 5). The same applies to the phenyl ring where the transition probabilities of $\chi^7$, $\chi^8$, and one valence angle are not taken into account; the number of variables for the Phe side chain is, therefore, 12. It turns out that the total number of angles considered is $K'$ = 92 − 24 = 68. The above discussion about the determination of the relevant set of angles is general and is expected to apply to any molecule. In fact, when $K'$ was increased ($K'$ > 68), the results for $\bar{F}^{\rm B}$(IS), the approximate free energy (eq 21), were found to become worse [i.e., $\bar{F}^{\rm B}(b = 1) > \bar{F}^{\rm B}(b = 0)$ ($\bar{F}^{\rm B}$(IS) should decrease as the approximation is improved)], because of the inclusion of these nonindependent variables. Also, one can argue that $\chi^2$ and $\theta$(C$^\beta$–C$^\gamma$–C$^\delta$) of proline are very much determined when the other two angles, ($\chi^1$ and $\theta$(C$^\alpha$–C$^\beta$–C$^\gamma$)) are specified and should be eliminated. However, the behavior of the results for $\bar{F}^{\rm B}$(IS) with and without these angles was found to be comparable. This applies also to two angles of the phenyl ring and to $\psi$ and $\theta$(C$^\alpha$–C′–N) of Phe$^6$ which close the backbone ring.

In the first stage of the calculation, one has to determine from the sample the ranges $\Delta\alpha_k$, $1 \leq k \leq K$, within which the dihedral and valence angles change. Table I shows these ranges for both the peptide in vacuum and in the crystal. We shall discuss here the vacuum results, whereas those for the crystal are discussed in section III.B. The calculations reveal that the valence angles

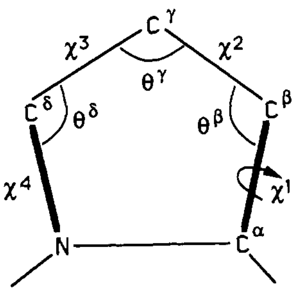

**Figure 5.** Construction of the proline side chain with the scanning procedure. The directions of the $C^{\alpha}$–$C^{\beta}$ and $C^{\delta}$–N bonds (shown with thick lines) are mostly determined by the backbone conformation. Therefore, the conformation of the ring can be determined by $\chi_1$, $\theta^{\beta}$, $\chi_2$, and $\theta^{\gamma}$, and these are the only angles treated as variables.

change in maximal ranges $\Delta\theta$, $15° \leq \Delta\theta \leq 35°$, whereas the corresponding standard deviations $\sigma_\theta$ around the average values of $\theta$ lie in a significantly smaller range, $1.77° \leq \sigma_\theta \leq 4.2°$. The same occurs for the phenyl ring where the ranges $\Delta\chi$ in which the dihedral angles $\chi$ fall are $\Delta\chi \sim 50°$ (see Table I) while the corresponding standard deviation value $\sigma_\chi$ is smaller, $\sigma_\chi \sim 6°$. As seen, the $\Delta\alpha_k$ values for the same angle in residues $j$ and $j + 3$ $(1 \leq j \leq 3)$ (e.g., $\Delta\phi$ of Ala$^1$ and Ala$^4$) are very close, differing by no more than 15° in the vacuum sample. The only significant difference, 230 versus 360, is observed for $\chi^1$ of Ala$^1$ and Ala$^4$, respectively (see discussion in refs 45 and 46). Also, the limits $\alpha_k$(max) and $\alpha_k$(min) of "symmetrically" related angles were found to be close to each other. This means that *on average* the two equal (in terms of sequence) halves of the molecule are also conformationally equal, which suggests that transition probabilities of symmetric local states can, to a good approximation, be considered identical [this will apply only when the values of the correlation parameter $b$ that are studied are relatively small, as in the present study $(b \leq 3)$; notice that for an exact LS method, at each step all the previous angles are considered and therefore the above symmetry cannot be used]. Thus, the data base for these probabilities doubles; this requires defining for symmetrically related angles (i.e., $\phi_j$ and $\phi_{j+3}$, $\psi_j$ and $\psi_{j+3}$, etc., $j$ = 1, 3, or $\alpha_k$ and $\alpha_{k+46}$, $k$ = 1, 46; see eq 31),

$$\Delta\alpha_k = \Delta\alpha_{k+46} = \max(\alpha_k,\alpha_{k+46}) - \min(\alpha_k,\alpha_{k+46}) \quad (38)$$

The lowest approximation for the entropy $S$ is based on $b$ = 0 (which means that correlations between successive angles are ignored) and on the assumption that the distribution of angles within the ranges $\Delta\alpha_k$ is homogeneous; i.e., the discretization parameter, $l$, is 1. The transition probability densities are (compare with eq 33 and see definition of notation for $b$ = 0 following eq 11)

$$\rho(\alpha_k,b{=}0,l{=}1) = 1/\Delta\alpha_k \quad (39)$$

and the approximate entropy (see eqs 16 and 20) is,

$$\bar{S}^{\mathrm{A}}(b{=}0,l{=}1) = -n^{-1}k_{\mathrm{B}}\sum_{k=1}^{K'} \ln \frac{1}{\Delta\alpha k} \quad (40)$$

where $k$ runs over the $K'$ = 68 backbone and side chain angles that are taken into account, $n$ is the sample size, and the bar above $S$ means estimation. Better approximations for $b$ = 0 are obtained by increasing $l$, the discretization parameter. In this case, the transition probability densities (see eq 33) are

$$\rho(\alpha_k,b{=}0,l) = n(\nu_k)/[n\Delta\alpha_{k,l}] \quad (41)$$

where the angle $\alpha_k$ belongs to segment $\nu_k$, $n(\nu_k)$ is the number of times $\nu_k$ appears in the sample, and $\Delta\alpha_{k,l}$ is the segment length (eq 32).

For $b \geq 1$ the transition probabilities are calculated in the following way. An order is defined among the backbone dihedral and valence angles $\phi$(Ala$^1$), $\theta$(N–$C^{\alpha}$–C′) (of Ala$^1$), $\psi$(Ala$^1$), ..., etc. and for each side chain (e.g., for proline $\chi$($C^{\alpha}$–$C^{\beta}$), $\theta$($C^{\alpha}$–$C^{\beta}$–$C^{\gamma}$), ..., etc.). Then local states are defined and two sets of transition probabilities are calculated from the sample for the backbone and side chains, respectively (see section II.C). Two types of approximations are defined for the side chains: (1) $l \geq 1$ and correlations between the angles are ignored (i.e., $b$ = 0); (2) $l > 1$ and the correlations based on $b$ = 1 are taken into account (see the discussions preceding eq 9 and before and after eq 13). The transition probability densities for the backbone are based on $b \leq 3$. Thus, for $b$ = 1 the backbone transition probability densities for Ala$^1$ are (for simplicity we omit, in most cases, the variables $b$ and $l$): $\rho(\phi, b = 0, l)$, $\rho(\theta(\mathrm{N–C^{\alpha}–C'})|\phi)$, $\rho(\psi|\theta(\mathrm{N–C^{\alpha}–C'}))$, $\rho(\theta(\mathrm{C^{\alpha}–C'–N})|\chi^1)$, $\rho(\omega|\theta(\mathrm{C^{\alpha}–C'–N}))$, $\rho(\theta(\mathrm{C'–N–C^{\alpha}}|\omega)$, and $\rho(\phi|\theta(\mathrm{C'–N–C^{\alpha}}))$, where in the last two $\rho$'s, N and $C^{\alpha}$ are of Pro$^1$. Notice that (1) the probability density for $\phi$(Ala$^1$), which is the first variable in the buildup procedure, is based on $b$ = 0 rather than $b$ = 1 and (2) $\theta(\mathrm{C^{\alpha}–C'–N})$ depends on $\chi^1$ (see eq 13). In the same way the probability densities are defined for the other backbone angles and for higher values of $b$ $(b \leq 3)$. Using $b$ = 1, the probability density for $\chi^1$ is $\rho(\chi^1|\psi)$ for all the amino acid residues (see eq 13). For proline we have found that the best approximations (i.e., which leads to maximal values of the approximate free energy $\bar{F}^{\mathrm{A}}$ and minimal values for $\bar{F}^{\mathrm{B}}$(IS), respectively) are obtained when the probability density of $\theta(\mathrm{C^{\alpha}–C^{\beta}–C^{\gamma}})$ and $\chi^2$ are based on $b$ = 0 (i.e., no correlations), while $\theta(\mathrm{C^{\beta}–C^{\gamma}–C^{\delta}})$ is correlated, i.e., $\rho(\theta(\mathrm{C^{\beta}–C^{\gamma}–C^{\delta}}|\chi^2)$ is used. For D-Phe we employ $\rho(\theta(\mathrm{C^{\alpha}–C^{\beta}–C^{\gamma}}|\chi^1)$, $\rho(\chi^2|\theta(\mathrm{C^{\alpha}–C^{\beta}–C^{\gamma}})$, and $\rho(\theta(\mathrm{C^{\beta}–C^{\gamma}–C^{\delta}})|\chi^2)$ (see Figure 4). It has been found that the best way to treat the phenyl ring is to define probability densities based on $b$ = 0 (no correlations) for $\chi^3$ and $\theta(\mathrm{C^{\gamma}–C^{\delta1}–C^{\epsilon1}})$, while the probability densities for the following angles (i.e., $\chi^4$, $\theta(\mathrm{C^{\delta1}–C^{\epsilon1}–C^{\zeta}})$, etc.) are based on $b$ = 1 (i.e., $\rho(\chi^4|\theta(\mathrm{C^{\gamma}–C^{\delta1}–C^{\epsilon1}}))$, etc.). These conclusions are general and can be applied to other side chains as well.

**1. Results for $F^{\mathrm{A}}$ and $\sigma_F^{\mathrm{A}}$.** Table II presents results for the approximate free energy $F^{\mathrm{A}}$ (eq 18) and its fluctuation $\sigma_F^{\mathrm{A}}$ (eq 29) [estimated by $\bar{F}^{\alpha}$ (eq 20) and $\bar{\sigma}_F^{\mathrm{A}}$ (eq 30)]. As mentioned above, in order to use these equations for a continuum model, the probabilities are replaced by probability densities (eq 34), and the summation (eqs 20 and 18) is carried out over the polypeptide sample rather than over the sample of SAWs. Eight different approximations defined by the correlation parameter $b$ are studied, in which the backbone probability density is described by $b$ = 0, 1, 2, and 3 and the side chains by $b$ = 0 and $b$ = 1 (i.e., approximations 1 and 2, respectively, discussed in the previous paragraph). Results for these eight approximations are obtained for various values of the discretization parameter $l$ from $l$ = 3 to $l$ = 40. For $F^{\mathrm{A}}$ we also present the result for the lowest approximation, based on $b$ = 0 and $l$ = 1 (see eqs 20 and 40). As expected, for each value of $b$, $\bar{F}^{\mathrm{A}}$ increases and $\bar{\sigma}_F^{\mathrm{A}}$ decreases as $l$ is increased, i.e., as the approximation improves. The same behavior is observed when $b$ is increased for a given $l$. Thus, $\bar{F}^{\mathrm{A}}$ increases from 23.47 kcal/mol per residue plus an additive constant for $b$ = 0 and $l$ = 1 (the worst approximation) to 29.45 (in the same units) for $l$ = 40 and $b$ = 2 for the backbone and $b$ = 1 for the side chains (the best approximation studied). The largest change in $\bar{F}^{\mathrm{A}}$ occurs in changing $l$ from 1 to 3. Correspondingly, the fluctuation $\bar{\sigma}_F^{\mathrm{A}}$ decreases from 1.003 kcal/mol per residue (for $l$ = 3 and $b$(back) = $b$(side) = 0) to 0.910 kcal/mol per residue (for $l$ = 40, $b$(back) = 2 and $b$(side) = 1).

It should be noted that all these results depend on the definition of the region $\Omega$ in phase space (eq 35), which in our case is defined by the sample and which constitutes a part of the region $\Delta\alpha_1 \times \Delta\alpha_2 \times ... \times \Delta\alpha_K$ defined by the $K$ = 92 internal angles. Therefore, for longer trajectories, both $\Omega$, and the $\Delta\alpha_k$'s, are expected to increase slightly, which will increase the entropy. [Notice that increasing the trajectory in general leads to a relatively small change in the average energy since the opposite energy fluctuations approximately cancel each other. On the other hand, entropy fluctuations are not canceled out, and therefore they always increase the average entropy; this can easily be deduced by considering a Gaussian distribution, for example.] Thus, estimating the accuracy of the results (for any given value of $b$ and $l$) is not straightforward, since it is affected by the size of $\Omega$, the statistical error of the transition probability densities, which depends on the

**Table II.** Results[a,b] for the Free Energy $\bar{F}^A$, Its Fluctuation $\bar{\sigma}_F{}^A$, and $\bar{F}^B$(IS) for the Peptide in Vacuum[c,d]

| $l$/$b$(back),$b$(side) | 0, 0 | 1, 0 | 2, 0 | 3, 0 | 0, 1 | 1, 1 | 2, 1 | 3, 1 |
|---|---|---|---|---|---|---|---|---|
| | | | | $\bar{F}^A$ | | | | |
| 1 | 23.47 | | | | | | | |
| 3 | 27.01 | 27.04 | 27.08 | 27.09 | 27.02 | 27.05 | 27.09 | 27.10 |
| 8 | 28.44 | 28.54 | 28.67 | 28.71 | 28.49 | 28.59 | 28.72 | 28.76 |
| 16 | 28.82 | 28.94 | 29.10 | 29.22 | 28.88 | 29.00 | 29.17 | 29.28 |
| 20 | 28.88 | 28.99 | 29.17 | | 28.94 | 29.06 | 29.24 | |
| 30 | 28.93 | 29.05 | 29.28 | | 29.00 | 29.12 | 29.35 | |
| 40 | 28.95 | 29.08 | 29.38 | | 29.02 | 29.15 | 29.45 | |
| | | | | $\bar{\sigma}_F{}^A$ | | | | |
| 3 | 1.003 | 0.993 | 0.982 | 0.979 | 0.996 | 0.987 | 0.977 | 0.975 |
| 8 | 0.978 | 0.963 | 0.950 | 0.946 | 0.964 | 0.951 | 0.942 | 0.938 |
| 16 | 0.974 | 0.955 | 0.938 | 0.924 | 0.959 | 0.942 | 0.930 | 0.918 |
| 20 | 0.972 | 0.952 | 0.933 | | 0.956 | 0.939 | 0.925 | |
| 30 | 0.972 | 0.950 | 0.925 | | 0.955 | 0.936 | 0.917 | |
| 40 | 0.971 | 0.948 | 0.917 | | 0.953 | 0.933 | 0.910 | |
| | | | | $\bar{F}^B$(IS) | | | | |
| 12 | 33.18 | 33.02 | | | 33.10 | 32.92 | | |
| 14 | 33.28 | 33.19 | | | 33.19 | 33.08 | | |
| 16 | 33.13 | 33.06 | | | 32.97 | 32.90 | | |
| 18 | 33.22 | 33.10 | | | 33.13 | 33.01 | | |
| 20 | 33.46 | 33.37 | | | 33.33 | 33.23 | | |
| 25 | 33.34 | 33.26 | | | 33.24 | 33.14 | | |
| 30 | 33.50 | 33.41 | | | 33.41 | 33.32 | | |
| 40 | 33.47 | 33.38 | | | 33.37 | 33.26 | | |

[a] $l$ is the discretization parameter; $b$(back) and $b$(side) define the approximation for the backbone and the side chains, respectively, where $b$ is the correlation parameter (for details see the text). [b] $\bar{F}^A$, $\bar{\sigma}_F{}^A$, and $\bar{F}^B$(IS) are defined in eqs 20, 30, and 24, respectively. [c] The results for $\bar{F}^A$ and $\bar{F}^B$(IS) are in kcal/mol per residue, plus an additive constant, which is the same for all cases; those for $\bar{\sigma}_F{}^A$ are in kcal/mol per residue. [d] The estimated statistical error increases as $l$ and $b$ are increased. The maximal estimated error of $\bar{F}^A$ is ±0.12, whereas that for $\bar{F}^B$(IS) is expected to be five times larger; the error of $\bar{\sigma}_F{}^A$ is ±0.005 (see discussion in text).

sample size, $n$, and, for a given set of transition probability densities also by the statistical error of the estimator $\bar{F}^A$ (eq 20) (which depends on $n$ and $\sigma_F{}^A$). We have, therefore, estimated, in Table II, the statistical errors from values of $\bar{F}^A$ and $\bar{\sigma}_F{}^A$ obtained for various sample sizes $n$. In these calculations the values of $\Delta\alpha_k$ (eq 38 and Table I) were kept fixed at those obtained for the entire sample of 112 500 conformations [see the caption of Table II]. The statistical error obtained in that way is, however, somewhat subjective. Also, the statistical errors do not represent independent uncertainties of the results; on the contrary, the values of $\bar{F}^A$ (and of $\bar{\sigma}_F{}^A$) change in a correlated manner for different sample sizes $n$. As mentioned in section II, for the exact free energy $F$ (eq 5) the fluctuation, $\sigma_F$, = 0, whereas the fluctuation of the energy and entropy, $\sigma_E$ and $\sigma_S$, respectively, increase as $N^{1/2}$ with increasing system size $N$. One therefore expects that for a good enough approximation, $\sigma_F{}^A$ (eq 29) would become smaller than $\sigma_E$ (eq 7). This was indeed observed in simulations of Ising models.[32,33] However, the value of the fluctuation obtained for the best approximation in the table, $\bar{\sigma}_F{}^A = 0.910(4)$ is still slightly larger than the energy fluctuation, $\bar{\sigma}_E = 0.8581(4)$ kcal/mol per residue. Obviously, one can define better approximations by increasing $b$ and $l$; however, the number of local states increases dramatically and thus very large samples are required in order to obtain an adequate data base.

**2. Results for $F^B$.** For a good enough approximation, the results for the approximate free energy $F^B$ (eqs 21, 24, and 28) are expected to constitute an upper bound of the correct free energy (see discussion following eq 22). It should be noted that it is more difficult to obtain a reliable estimate for $F^B$ than for $F^A$. This is because the largest contribution to $F^A$ comes from the most probable conformations with respect to the sampling probability density (i.e., the Boltzmann distribution) (see eq 36). These conformations have typical equilibrium energies [which are close to the ensemble average value, $E$ (eq 3)] and are also expected to consist of typical local states; therefore, the data base for the corresponding transition probabilities is sufficiently large even for relatively large values of $l$ and $b$. $F^B$, on the other hand, is defined with the approximate $\rho'([\alpha_k], b, l)$ (based on eqs 15 and 34) but has to be estimated from a sample generated with the correct Boltzmann probability density $\rho([\alpha_K])$ (eq 36). Therefore, as discussed above (section II.D.1), if these probability distributions are different, the typical conformations with respect to $\rho'$ will belong to the "tail" of $\rho([\alpha_K])$ (see Figure 2). Therefore, only a small number of conformations will contribute significantly to $F^B$. One would therefore expect that $F^B$ would be statistically less reliable than $F^A$.

A measure of the effective sample size for $F^B$ is provided by the number of accepted conformations, $n_{\text{accept}}$ (eq 27), obtained with the Schmidt procedure. In fact, the values of $n_{\text{accept}}$ obtained (they depend somewhat, of course, on the sequence of random numbers) have never exceeded 70 (out of a total sample of 112 500), which means that the effective sample for $F^B$ is very small. This should be compared with $100 \le n_{\text{accept}} \le 1200$ obtained for the $\alpha$-helical state of decaglycine in ref 34. The present smaller values of $n_{\text{accept}}$ stem from the fact that better approximations (larger values of $b$ and $l$) are required to treat a cyclic molecule than a linear one since the internal coordinates must be highly correlated in order to achieve ring closure. The existence of side chains and the use of flexible geometry also add to the difficulty of handling cyclo-(Ala-Pro-D-Phe)$_2$ as compared to decaglycine.

The results for $\bar{F}^B$(IS) in Table II reveal that for each pair of values $b$(back) and $b$(side), the results for $\bar{F}^B$(IS) (eq 24) do not decrease with increasing $l$, i.e., improving the approximation, but are smaller for $12 \le l \le 18$ than for $l > 18$. This probably stems from the fact that the data base becomes insufficient for the larger values of $l$. On the other hand, for each $l$, $\bar{F}^B$(IS) decreases with improvement in the approximation. Also, as expected, the results for the approximations (0, 1) and (1, 1) are always smaller than those for (0, 0) and (1, 0), respectively. This behavior has been found to occur for sample sizes of $n$ = 90 000 and up and, therefore, in spite of the relatively large statistical error, we consider it to be significant. However, for $l < 12$ an opposite trend has been observed. It should be pointed out that the values for $\bar{F}^B$(A) (eq 28) (i.e., those calculated with the Schmidt procedure) in most cases increased for a given $l$ ($l \ge 12$) rather than decreased with improving the approximation. However, these values are statistically less reliable than those of $\bar{F}^B$(IS) since the Schmidt procedure is based on an additional stochastic process.

The results for $\bar{F}^B$(IS) are not accurate enough to define approximations for the average free energy, $F^M$ (eq 23), that converge satisfactorily to $F$ (sufficiently accurate approximations were obtained for the $\alpha$-helical state of decaglycine in ref 34), and, indeed, the present approximations are significantly worse than those obtained for the $\alpha$-helix. The two molecules have comparable numbers of atoms, but the difference, $F^A - F^B$, for the best approximations is ~2.03 kcal/mol for the $\alpha$-helix versus ~21.3 kcal/mol for cyclo-(Ala-Pro-D-Phe)$_2$. However, in order to get some feeling for the accuracy of the present approximations, we have calculated for $l = 20$ and ($b$(back), $b$(side)) = (2, 1) the value $F^M = 31.24$ (eq 23) which should be compared to $F^A = 29.24$ and $F^B = 33.23$, all in kcal/mol per residue (plus the same additive constant). The corresponding values of the entropy are $TS^A = 21.29$ and $TS^M = 19.22$ in the same units, where $T = 306.6$ K is the average temperature during the simulation.

In summary, the behavior of the results for the approximate free energies $F^A$ and $F^B$ and the fluctuation of $F^A$, $\sigma_F{}^A$, for the cyclic polypeptide in vacuum has shown that the method is reliable (i.e., certain inequality relations are satisfied). However, in order to obtain converging results for $F^M$, larger samples are required, which would increase the number of accepted conformations leading, thereby, to better approximations for $F^B$. However, we have decided to avoid these extra calculations (which are feasible) since $F^B$ is not defined for the peptide in the crystal, and hence $S^A$ (which is obtained with sufficient accuracy) is used to compare the entropies of the peptide in the two environments.

**B. The Crystal Environment.** The unit cell contains two peptides and 16 water molecules. In order to apply the LS method, one should first envisage an appropriate exact scanning procedure. One can start building peptide 1, for example (in an empty cell), step-by-step as previously described. However, in this case the future scanning is carried out over all the peptide angles and also over the various configurations of the waters and peptide 2; these configurations limit the conformational freedom of chain 1. Thus, if the cell construction is stopped after building peptide 1, the probability of construction defines an entropy of peptide 1 in the unit cell. This entropy, denoted $S_p$(cell), is the entropy of a molecule with potential of mean force[7,8] (see discussion following eq 42 below). If construction of the cell is continued, one can obtain the entropy of the whole unit cell. In this paper, however, we limit ourselves to the study of the peptide entropy. Thus, the LS method for the peptide in the crystal is carried out in basically the same way as described for the peptide in vacuum (ignoring, thereby, the translational entropy part of $S_p$(cell) which is, however, expected to be small); the translational entropy of the molecule in vacuum also has not been taken into account. Here, however, because the two peptide chains are positioned symmetrically in the cell (i.e., they have similar environments), one can, to a good approximation, treat them as independent structures, thereby doubling the sample size ($n$ = 72 500 rather than 36 250).

Let us now develop a formal expression for the entropy $S_p$(cell) (discussed above) of a polypeptide chain in the unit cell. For the sake of simplicity we shall use the notation of a discrete model. The cell partition function, $Z$(cell), can be written as

$$Z(\text{cell}) = \sum_{i \in \Omega} \exp[-F_i(\text{total})/k_B T] \qquad (42)$$

where $i$ runs over all the conformations of peptide 1 in the ensemble (phase space) $\Omega$, for example, and $F_i$(total) is a free energy function that is obtained from a partition function (eq 5) which is based on a summation over all the configurations of the water molecules and peptide 2 for a given conformation $i$ of peptide 1 (see refs 7 and 8). The ensemble probability of $i$ in the cell is thus,

$$P_i{}^B(\text{cell}) = \exp[-F_i(\text{total})/k_B T]/Z(\text{cell}) \qquad (43)$$

and the conformational entropy of peptide 1 in the cell, $S_p$(cell), is

$$S_p(\text{cell}) = -k_B \sum_{i \in \Omega} P_i{}^B(\text{cell}) \ln P_i{}^B(\text{cell}) \qquad (44)$$

$S_p$(cell), defined above, is a measure of the conformational freedom of peptide 1 in the cell and can be compared with the entropy of the peptide in vacuum. Notice that the free energy of the whole cell, $F$(total), can be expressed in terms of the probability $P_i{}^B$(cell) and the free energy $F_i$(total)

$$F(\text{total}) = -kT \ln Z(\text{total}) = \sum_{i \in \Omega} F_i(\text{total}) + kT \ln P_i{}^B(\text{cell}) \qquad (45)$$

where the fluctuation of $F$(total) can be shown to vanish.[53] As for the peptide in vacuum, one can define an approximation $S_p{}^A$(cell) [based on an approximate probability distribution $P_i(b)$ (see eq 16)] and in principle also lower and upper bounds for $F$(total), $F^A$(total), and $F^B$(total), respectively. However, calculation of the two latter quantities is not practical since $F_i$(total) is unknown. On the other hand, one cannot define a partition function solely for peptide 1 in the cell from which free energy, energy, and entropy can be derived from the usual thermodynamic equations. However, it is still useful to define several energies which are related to peptide 1 in the cell. Thus, we define the intramolecular energy $E_i$(intra) which consists of the intrapeptide interactions for conformation $i$. Another energy is $E_i$(peptide), which include $E_i$(intra) as well as the ensemble average (over all configurations of the water molecules and peptide 2) of the interaction energy of the atoms in the unit cell (those of the 16 waters and peptide 2) with the atoms of peptide 1 (fixed in conformation $i$). One can also define $E_i$(total) which is the ensemble average of the total energy of the unit cell for conformation $i$ of peptide 1. The same can be defined for peptide 2. The average of $E_i$(intra) over the peptide configuration $i$ (denoted $E$(intra)) can be compared to the average energy of the peptide in vacuum; the difference between these energies measures the strain energy, which results from the forces exerted on the peptide chain by the crystal environment.[45] $E_i$(total) constitutes part of $F_i$(total), and therefore its ensemble average (over $i$), $E$(total) together with the approximate entropy $S_p{}^A$(cell) provides an approximation for $F$(total) (using the relation $F = E - TS$). The same relation enables one to define free energy function of peptide 1 which is based on $S_p{}^A$(cell) and the ensemble average (over $i$) of $E_i$(peptide) (denoted $E$(peptide)). This free energy can be used as an approximate measure of binding of a ligand to a protein (see, for example, ref 3). The fluctuations of these three approximate free energies are not zero; however, their magnitudes are important since they determine the statistical error of the corresponding free energies.

The values of the ranges of the internal coordinates $\Delta\alpha_k$ (eq 38) for the polypeptide in the crystal are presented in the lower part of Table I. The values of $\Delta\alpha_k$ for the backbone and many of the side chain dihedral angles are significantly smaller in the crystal than in vacuum. On the other hand, the valence angles $\theta$ in the two environments are close. This suggests that the peptide has lower entropy in the crystal than in vacuum. The table also shows that the values of $\Delta\alpha_k$ for corresponding angles of residue $j$ and $j + 3$ ($j$ = 1, 3) are close. However, the differences between these pairs of values are slightly larger in the crystal than in vacuum. This is probably both because of the smaller sample size of the former model and the fact that the environments of residues $j$ and $j + 3$ (e.g., Ala$^1$ and Ala$^4$) were slightly asymmetric at the beginning of the simulation (since the water molecules were not symmetrically arranged in the initial system) and remain so throughout the simulation.[45]

In Table III results for the entropy $S_p{}^A$(cell) [estimated by $\bar{S}_p{}^A$(cell)] of the peptide in the crystal are presented for various approximations $b$ and $l$. As expected (eq 17), $S_p{}^A$(cell) decreases as $b$(back) and $l$ are increased. This means that the three approximate free energies $F^A$ (eq 18) (based on $S_p{}^A$(cell) and the ensemble averages (over $i$) $E$(intra), $E$(peptide), and $E$(total)) should all increase as $S_p{}^A$(cell) decreases. From the sample we estimate the values $E$(intra) = 306.4 (3), $E$(peptide) = 219.4 (2), and $E$(total) = 322.4 (3) (all in kcal/mol). In these calculations, the energies for each $i$, $E_i$(peptide) and $E_i$(total), defined above as ensemble averages, were estimated by the energies of the instantaneous configuration of the system associated with each $i$. $E$(intra) is slightly larger than the vacuum energy,[45] 303.17 (2) kcal/mol. One can also calculate the fluctuations $\sigma_E$ (eq 7) of the above three energies and the corresponding free energy fluctuations $\sigma_F{}^A$ (eqs 29 and 30). We have found that the values of $\bar{\sigma}_F{}^A$ for the three definitions of the energy are comparable to the corresponding energy fluctuations and they always decrease

**Table III.** Results for the Approximate Entropy $\bar{S}_p^A$ for the Peptide in the Crystal[a-c]

| | $b$(back),$b$(side) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $l$ | 0, 0 | 1, 0 | 2, 0 | 3, 0 | 0, 1 | 1, 1 | 2, 1 | 3, 1 |
| 1 | 0.0830 | | | | | | | |
| 3 | 0.0700 | 0.0699 | 0.0699 | 0.0698 | 0.0700 | 0.0699 | 0.0698 | 0.0698 |
| 8 | 0.0663 | 0.0662 | 0.0659 | 0.0657 | 0.0662 | 0.0661 | 0.0658 | 0.0656 |
| 20 | 0.0651 | 0.0650 | 0.0645 | | 0.0650 | 0.0649 | 0.0644 | |
| 30 | 0.0650 | 0.0648 | 0.0642 | | 0.0649 | 0.0647 | 0.0640 | |
| 40 | 0.0649 | 0.0648 | 0.0638 | | 0.0648 | 0.0646 | 0.0636 | |

[a] $l$ is the discretization parameter; $b$(back) and $b$(side) define the approximation for the backbone and the side chains, respectively, where $b$ is the correlation parameter (for details see the text). [b] $\bar{S}_p^A$ is defined in eqs 16 and 20 (see also eq 44). [c] The results are in kcal/(mol deg) per residue plus an additive constant; the error is ±0.0003. For comparison, for $l = 40$ and $[b(\text{back}),b(\text{side})] = [2,1]$, $\bar{S}^A = 0.0689$ in vacuum.

**Table IV.** Results for $T\Delta S^A$, the Difference between the Peptide Entropy in Vacuum and in the Crystal[a,b]

| | $b$(back),$b$(side) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $l$ | 0, 0 | 1, 0 | 2, 0 | 3, 0 | 0, 1 | 1, 1 | 2, 1 | 3, 1 |
| 1 | 10.01 | | | | | | | |
| 6 | 11.01 | 10.70 | 10.47 | 10.51 | 10.84 | 10.55 | 10.33 | 10.36 |
| 8 | 10.59 | 10.23 | 9.98 | 10.07 | 10.42 | 10.07 | 9.81 | 9.90 |
| 12 | 10.42 | 10.03 | 9.75 | 9.90 | 10.23 | 9.83 | 9.53 | 9.72 |
| 14 | 10.25 | 9.85 | 9.57 | 9.77 | 10.07 | 9.66 | 9.37 | 9.59 |
| 16 | 10.22 | 9.79 | 9.51 | 9.81 | 10.01 | 9.59 | 9.31 | 9.61 |
| 20 | 10.16 | 9.74 | 9.46 | | 9.94 | 9.50 | 9.26 | |
| 30 | 10.05 | 9.61 | 9.50 | | 9.85 | 9.42 | 9.27 | |
| 40 | 10.05 | 9.61 | 9.66 | | 9.83 | 9.42 | 9.46 | |

[a] $\Delta S^A$ is defined in eq 46. $l$ is the discretization parameter; $b$(back) and $b$(side) define the approximation for the backbone and the side chains respectively, where $b$ is the correlation parameter (for details see the text). $T = 307.9$ K. [b] $T\Delta S^A$ is given in kcal/mol. The final estimate of $\Delta S^A$ has been obtained by averaging the results within the dashed square.

as the approximation improves; this is important since $\bar{\sigma}_F{}^A$ determines the statistical error. Our results are $\bar{\sigma}_F{}^A = 1.055$ versus $\bar{\sigma}_E = 1.069$ (intra), 1.120 versus 1.138 (peptide), and 1.434 versus 1.390 (total) (all in kcal/mol per residue) for the best approximation, $l = 40$, $b = 2$ (backbone), and $b = 1$ (side chains). As has been mentioned, these energies and free energies are of interest since their averages can be used as approximate criteria for determining protein–ligand binding.

**1. LS Results for $\Delta S^A$.** As discussed above, the free energy of the *peptide* in the cell is not well defined (only the free energy of the whole cell is well defined thermodynamically), and therefore only its entropy $S_p$(cell) (eq 42) can be compared with that of the peptide in vacuum. Since the average entropy $S^M = (E - F^M)/T$ (see eqs 5 and 23) is not defined for the peptide in the crystal, one can estimate only $S_p{}^A$(cell) (see eq 16) which, for simplicity, will be denoted $S_p{}^A$ (we denote by $S^A{}_{\text{vac}}$ the approximate entropy of the peptide in vacuum). In Table IV, values of $\Delta S^A$, where

$$\Delta S^A(b, l) = S^A{}_{\text{vac}}(b, l) - S_p{}^A(b, l) \qquad (46)$$

are given for the discretization parameter $l$ = 1, 6, 8, 12, 14, 16, 20, 30, and 40 using correlation parameter $b$(back) = 0, 1, 2, and 3 for the backbone and $b$(side) = 0 and 1 for the side chains. In this calculation the temperature $T$ = 307.9 K is the average value of 306.6 K and 309.27 K, which are the average temperatures obtained from the vacuum and the crystal simulations, respectively. For each pair [$b$(back), $b$(side)] the results for $\Delta S^A$ decrease as $l$ is increased. This shows that $S^A$ decreases more rapidly in the vacuum system as $l$ is increased, indicating that larger values of $l$ are required for the vacuum system. This reflects the larger ranges of dihedral and valence angles, $\Delta\alpha_k$, for the vacuum case. However, for the two largest values of $l$ in each column, this decrease actually ends and in some cases (for $b$(back) = 2 and 3) $\Delta S^A$ even increases, which means that, within the statistical error, convergence has been attained. Correspondingly, for each value of $l$ and $b$(side), the results for $b$(back) decrease in going from $b$(back) = 0 to 2 (except for $l$ = 40), indicating that $S^A$ for the vacuum case decreases more strongly than for the crystal as $b$(back) is increased. This probably reflects the fact that the angle–angle correlations along the chain backbone are slightly stronger for the peptide in vacuum than in the crystal (because for the peptide in vacuum the ring closure condition must be satisfied for larger values of the backbone ranges $\Delta\alpha_k$ than for the peptide in the crystal), and therefore larger values of $b$(back) are required to treat the vacuum system than are required to treat the crystal peptide equivalently. This effect, however, weakens with increasing $b$(back) where the difference $T\Delta S^A[b(\text{back})] - T\Delta S^A[b(\text{back}) + 1]$ is ~0.4–0.5 kcal/mol for $b$(back) = 0 and ~0.1–0.3 kcal/mol for $b$(back) = 1. However, for $b$(back) = 2, this tendency reverses where $\Delta S^A[b(\text{back}) = 3] > \Delta S^A[b(\text{back}) = 2]$. Again, one could interpret this behavior as a sign of convergence of $\Delta S^A(b,l)$. The results for $b$(side) = 0 (the left side of Table IV) are larger by ~0.2 kcal/mol than the corresponding values of $b$(side) = 1. This difference is relatively small, and it stems only from the larger decrease in the results for $S^A{}_{\text{vac}}$ than those for $S_p{}^A$ in going from $b$(side) = 0 to 1 (see results for $F^A$ and $S_p{}^A$ in Tables II and III, respectively). We have also calculated $\Delta S^A$ for the same backbone transition probabilities described above (i.e., based on $b$(back) ≤ 3) but for side chain transition probabilities that are better and worse than those that consist of $b$(side) = 0 and $b$(side) = 1, respectively. Thus, for the side chain dihedral angle $\chi^1$ we have used $\rho(\chi^1|\psi)$ (i.e., $b$(side) = 1) while $\rho(b = 0, l)$ has been defined for the other side chain angles. The results for $\Delta S^A$ obtained by this approximation are larger by not more than 0.6% than the corresponding values in Table IV based on $b$(side) = 1. This suggests that convergence of $\Delta S^A$ has been attained also with respect to the side chain contributions. It should also be pointed out that the $\Delta\alpha_k$ values for the dihedral and valence angles of the D-Phe rings are relatively small and comparable for the peptide in the two environments, and therefore their contribution to the correct $\Delta S$ is expected to be small. In fact, we have calculated a similar table to Table IV but without taking into account the angles of these rings, and the results were found to differ from those in Table IV by not more than 1%.

The convergence of the results of $\Delta S^A$ enables one to estimate $\Delta S$ by $\Delta S^A$. For that we average the results in Table IV for $b$(side) = 1, $l$ ≥ 16, and $b$(back) ≥ 1, obtaining

$$T\Delta S = T\Delta S^A = 9.4 \pm 0.8 \text{ kcal/mol}$$

The statistical error was obtained by calculating tables like Table IV for a sample of $n$ = 26 000 and $n$ = 80 000 for the crystal and vacuum simulations, respectively. Notice that the above statistical error does not stem from scatter in the results for $S_p{}^A$ and $S^A{}_{\text{vac}}$ for $l$ = 16–40; in fact, this scatter is much larger, ranging from 6.65 to 12.93 for $\min[S^A{}_{\text{vac}}] - \max[S_p{}^A]$ and $\max[S^A{}_{\text{vac}}]$ − min-

[$S_p^A$], respectively. It should be pointed out that the result $T\Delta S^A = 10.01$ kcal/mol for the worst approximation, $b$(back) = 0, $b$(side) = 0, and $l = 1$ (i.e., that which assumes homogeneous distribution) lies well within the statistical error of our estimate of 9.4 kcal/mol.

The fact that $\Delta S^A$, the difference of the approximate entropies, converges to the exact entropy difference, $\Delta S$, for relatively small values of the correlation parameter $b$(back), $b$(back) = 2 or 3, while the corresponding values of the entropies $S^A{}_{\text{vac}}$ and $S_p{}^A$ significantly overestimate (see end of section III.A.2) the correct values of the entropies $S$ (eq 4) and $S_p$(cell) (eq 44), respectively, means that the deviations $S^A{}_{\text{vac}} - S$ and $S_p{}^A - S_p$(cell) are comparable, and therefore they are cancelled out in $\Delta S^A$. As can be seen from Table I, the values of the ranges $\Delta\alpha_k$ are somewhat larger for many angles in the vacuum system than for the corresponding angles in the crystal case. We might expect, however, that in some cases the contribution to the difference in entropy from these angles would not be dependent on $l$. To explain this convergence, with respect to $l$, we consider the following example. Let $\Delta\alpha_k$(v) and $\Delta\alpha_k$(c) denote the ranges (eq 31) of the same angle $\alpha_k$ for the molecule in the vacuum and the crystal, respectively, where $\Delta\alpha_k(\text{v}) = a\Delta\alpha_k(\text{c})$ ($a > 1$); thus for the lowest approximation ($b = 0$, $l = 1$), one obtains $\Delta S^A(\alpha_k) = \ln a$ (see eq 40). For a better approximation ($b = 0$, $l$), one has $l$ transition probability densities $p(\nu_k, b = 0)/\Delta\alpha_{k,l}$ [$=\rho(\nu_k,b{=}0)$], $1 \leq \nu_k \leq l$ (see eq 33). If $p(\nu_k,b{=}0)$ is close to a Gaussian for both the peptide in vacuum and in the crystal, and because $\sum_k p(\nu_k,b{=}0) = 1$, the values of $p$ for segments $\nu_k$, $1 \leq \nu_k \leq l$ will be approximately the same in the two environments; therefore again

$$\Delta S^A(\alpha_k) \simeq \sum_{\nu_k=1}^{l} \{\ln [p(\nu_k)l/a\Delta\alpha_k(\text{c})] - \ln [p(\nu_k)l/\Delta\alpha_k(\text{c})]\} = \ln a \quad (47)$$

where $p(\nu_k,b{=}0)$ is denoted by $p(\nu_k)$ and $\Delta\alpha_{k,l} = \Delta\alpha_k/l$ (see eq 32). Thus if the angles were uncorrelated, $\Delta S^A$, that is based on $b = 0$ and a small value of $l$, would be a good approximation for $\Delta S$. However, the above assumption that the probability distributions $p(\nu_k,b{=}0)$ in the two environments are approximately equal might be wrong in a case, for example, where $\Delta\alpha_k$(c) is small (50–100°) while $\Delta\alpha_k(\text{v}) = 360°$. Also, for small values of $b$, the relation $\Delta S^A(b, l) \simeq \Delta S$ that is found in the present case might be incorrect if the range of correlations between angles (along the backbone) is *significantly* different for the two states being compared (see previous discussion in this section). In such a case, the values of $S^A(b,l)$ for the state with the shorter correlation range (denoted by $b_0$) are not expected to change as $b$ is increased beyond $b_0$, while the results for $S^A(b,l)$ of the second state will continue to decrease for $b > b_0$. Such an example was studied by Karplus and Kushick,[14] who used the quasiharmonic approximation, where the helical state of decaglycine was found to be more correlated than the extended state. In general, these correlations will depend on the potential energy function, the temperature, and the conformational state in which the molecule is located. However, it is difficult to predict the range of these correlations, and, therefore, one should verify for each molecule studied that the results for the approximation $\Delta S^A$ indeed converge. In fact, we reexamined the results in ref 34 for decaglycine for the difference in the approximate free energy $\Delta F^A = F^A(\text{helix}) - F^A(\text{hairpin})$ for the approximations $l$ = 4, 8, 12, and 16 and found for $b = 1$ the values, $\Delta F^A$ = 0.38, 0.38, 0.39, and 0.37 and for $b = 2$, $\Delta F^A$ = 0.38, 0.39, 0.39, and 0.39 respectively (all in kcal/mol). These results are equal within the statistical error to the value $\Delta F^M = F^M(\text{hairpin}) - F^M(\text{helix}) = 0.40(7)$ kcal/mol (see eq 23) which is considered to be the exact free energy difference within the statistical error; $\Delta F^A(b{=}0,l{=}1)$, however, was found to be slightly smaller, ~0.33 kcal/mol, but still within the statistical error of $\Delta F^M$. This is therefore another example where $\Delta F^A$ (and therefore $\Delta S^A$) constitutes an excellent approximation of $\Delta F$ ($\Delta S$).

We should point out again that the worst approximation $\Delta S^A(b{=}0,l{=}1)$ is expected, in cases where comparable ranges of correlations exist in the two states and the corresponding values of $\Delta\alpha_k$ are not extremely different, to be a good approximation for $\Delta S$. This is important since the number of local states in this approximation is the smallest, $K'$ (which is the number of angles $\alpha_k$ taken into account), and therefore the method can be applied efficiently to a protein of any size. Better approximations consist of $K'l^{b+1}$ local states and require larger computer memory and longer trajectories for generating a suitable data base.

**2. Results for $\Delta S^A$ Using the Harmonic Approximation.** We have also calculated the entropy using the harmonic approximation, i.e., from the normal mode eigenvalues (frequencies). For the peptide in vacuum, the lowest energy conformation found by minimizing structures sampled at regular intervals during the dynamics simulation was taken, and the normal modes were calculated around this lowest energy structure. For the crystal the energy of the complete unit cell was minimized, starting from the X-ray structure, and the normal mode analysis was carried out for peptide 1 and peptide 2 independently. In this latter analysis, the interactions between the peptides and their environment (to a cutoff of 15 Å) were taken into account. The entropy for the vacuum and crystal peptides was then calculated by the Einstein quantum mechanical harmonic oscillators formula[51] where,

$$TS = \sum_j \left[ \frac{h\nu_j}{\exp(h\nu_j/k_BT) - 1} - \ln\left(1 - \exp\left(-\frac{h\nu_j}{k_BT}\right)\right)\right] \quad (48)$$

and $j$ runs over all the vibrational frequencies $\nu_j$ and $h$ is Planck's constant. We also used the formula for classical oscillators where the difference in entropy between the peptides in the two environments is[14]

$$T\Delta S = k_B \ln [\textstyle\prod_j \nu_j(\text{c})/\prod_j \nu_j(\text{v})] \quad (49)$$

Here $\nu_j$(c) and $\nu_j$(v) are the frequencies of the peptide in the crystal and vacuum, respectively.

The lowest frequency of the peptide in vacuum is 9.5 cm$^{-1}$, while those for the two peptides in the crystal are 44.7 and 47.5 cm$^{-1}$. It should be noted that, for the molecule in vacuum, the six frequencies related to translation and rotation are zero; therefore, the entropy is based only on the $3N - 6$ (=258) frequencies of the vibrations. For the peptides in the crystal, however, we have found that all $3N$ frequencies are nonzero and that the normal modes corresponding to the 20 and 40 lowest frequencies (out of a total of $3N$ for peptide 1 and 2, respectively) include substantial nonzero components of translation and rotation, respectively. Translational movement was determined by a change in the center of mass upon displacement along each normal mode, and rotational movement was determined by calculation of the total angular momentum, the sum of the cross products $R_k \times m_k r_k$ for all atoms of the peptide, where $R_k$ is the vector from the center of mass of the peptide to atom $k$, $m_k$ is the mass, and $r_k$ is the displacement of atom $k$ along the normal mode (minus any translational movement). It should be pointed out that these numbers (20 and 40) depend on a cutoff for the contribution from rotation and translation that is somewhat arbitrary. Since the six zero frequencies corresponding to rotation and translation were eliminated from the calculation of the entropy of the isolated peptide, we have discarded the contribution of six frequencies of the crystal peptides in order to compare the entropies of the peptides in the two environments. These six frequencies were selected according to four different criteria: (1) the six lowest frequencies; (2) and (3) those with the largest translational and rotational components, respectively (these two sets of frequencies are not necessarily the same, nor are they necessarily the six lowest frequencies); and (4) the three frequencies with the largest translational components and the three with the largest rotational components. Differences in entropy, $\Delta S$ [$=S(\text{vacuum}) - S(\text{crystal})$], for peptide 1 were calculated by eliminating the contribution of each one of the above four groups of frequencies. The four results for $\Delta S$ were then averaged and the statistical error was obtained from the corresponding variance. The same was done for peptide 2. The results (in kcal/mol) are

$$T\Delta S(1, \text{quantum}) = 12.2 \pm 0.4$$

$$T\Delta S(1, \text{classical}) = 13.3 \pm 0.4$$

$$T\Delta S(2, \text{quantum}) = 11.3 \pm 0.4$$

$$T\Delta S(2, \text{classical}) = 12.4 \pm 0.4$$

where 1 and 2 denote peptides 1 and 2, respectively, and $T = 307.9$ K; the quantum and classical refer to results obtained by eqs 48 and 49, respectively. These results show a difference of ~1 kcal/mol between the corresponding results for peptides 1 and 2. The values of $\Delta S$(classical) are always larger by ~1 kcal/mol than the corresponding quantum mechanical values. The above results, as a whole, are 2–4 kcal/mol larger than the value of 9.4 obtained by the local states method. It should be pointed out again that our elimination of the contribution of only six frequencies for the peptides in the crystal is somewhat arbitrary since nonzero components of translation and rotation were found to exist for other frequencies as well (notice that elimination of additional frequencies would lower the values of $\Delta S$).

For the peptide in vacuum we have calculated the entropy for several other minimum energy conformations accessible to the molecule. The difference between the maximum and minimum values found for the quantum mechanical entropy ($TS$, eq 48) was 1.5 kcal/mol. Also, in the calculations described above of $S_p{}^A$(cell) for the peptide in the crystal, the data from both peptides were used together to obtain a single value of the entropy (Table IV). We have calculated tables similar to Tables I and IV for the peptides 1 and 2 separately. The corresponding values of $\Delta\alpha_k$ are obviously smaller than those based on the two peptides together, and thus the values of $T\Delta S^A$ are up to 1 kcal/mol larger than those of Table IV, but they are still smaller than those obtained with the harmonic approximations.

## IV. Summary

The LS method is general in the sense that it is not limited to handling the entropy of harmonic or quasiharmonic conformational changes but can be applied to any chain flexibility (see ref 37). The method provides approximations for the entropy which can be systematiclly improved by increasing the correlation and discretization parameters, $b$ and $l$, respectively.

Here the LS method was applied to samples obtained from molecular dynamics simulations of cyclo-(Ala-Pro-D-Phe)$_2$ in vacuum and in the crystal. The free energy functionals $F^B$ (eq 21) and $F^A$ (eq 18) and the fluctuation $\sigma_F{}^A$ (eq 29) of the latter were calculated ($F^B$ was only calculated for the peptide in vacuum) and shown to satisfy certain theoretical relations, which suggests that the method is reliable. In the usual application of the LS method to a chain in vacuum, one would seek to estimate $F^M = (F^A + F^B)/2$, which is expected to provide the best approximation for the correct free energy, $F$. However, this might require generating relatively large samples in order to adequately estimate $F^B$. Also, for a peptide in solvent or in the crystal, $F^B$ for the peptide chain is not well defined. Therefore, in such cases one can only calculate $S^A$ (or some approximations $F^A$) and the difference $\Delta S^A$ ($\Delta F^A$) between the $S^A$ ($F^A$) values of two different conformational states. An important conclusion is that in certain cases the results for $\Delta F^A(b,l)$ and $\Delta S^A(b,l)$ converge rapidly at small values of $b$ and $l$, which suggests that these functionals provide very good approximations for the correct $\Delta F$ and $\Delta S$, respectively. The computational advantage of using $F^A$ and $S^A$ rather than $F^B$ lies in the fact that they can be estimated efficiently from relatively small samples. Furthermore, if the ranges of the angle–angle correlations (measured by $b$) of two states are comparable and the corresponding values of $\Delta\alpha_k$ are not extremely different, the worst approximation, which ignores correlations and assumes homogeneous distribution of angles ($b = 0$, $l = 1$), provides a reasonable estimate of the correct $\Delta F$ (and $\Delta S$). This is also important since the above approximation requires the minimal number of local states $K'$ (i.e., the total number of angles), and therefore relatively small samples and very little computer memory are needed; thus, it can be applied efficiently to proteins of any size.

The LS method could be used (together with the appropriate energy contributions) to compare *approximately* the free energy of binding of various ligands to a protein, where at this stage the contribution of solvent entropy is ignored.[3] However, in principle, the entropy of diffusive systems such as a fluid can also be calculated with the LS method, since it was originally developed for Ising and lattice gas models.[31-33] This is planned for future work.

**Acknowledgment.** We thank Dr. Rick Fine for helpful discussions. This work was supported in part by grants from the National Institutes of Health (Grants GM30564 and GM45651) by the Supercomputer Computations Research Institute which is supported in part by the U.S. Department of Energy through Contract No. DE-FC05-85ER250000. Computer time was provided by the San Diego Supercomputer Center (SDSC) on the Cray X-MP/48 and Y-MP8/864.

**Registry No.** Cyclo-(Ala-Pro-D-PHE)$_2$, 52634-26-1.

# Ab Initio Studies of Fundamental Cluster Rearrangement Mechanisms

**David J. Wales* and Richard G. A. Bone**

*Contribution from the University Chemical Laboratories, Lensfield Road, Cambridge CB2 1EW, UK.* *Received December 16, 1991*

**Abstract:** Two fundamental cluster rearrangement mechanisms are investigated by ab initio calculations, namely the single diamond-square-diamond (DSD) process in $B_8H_8^{2-}$ and the square-diamond-diamond-square (SDDS) process in $C_5H_5^+$. Geometry optimizations and frequencies are compared for the SCF and second-order Møller–Plesset (MP2) approximations with basis sets ranging in size from STO-3G to double-$\zeta$ plus polarization (DZP). The results are in good agreement with expectations, especially orbital symmetry selection rules, and enable us to compare the effects of electron correlation and polarization functions upon the energy and the character of the stationary points. The topology of the potential energy surface of $C_5H_5^+$ is studied in detail, demonstrating that the SDDS mechanism allows all versions of the structure to be accessed.

## I. Introduction

In 1966 Lipscomb first proposed the diamond-square-diamond (DSD) process to account for rearrangements in boranes and carboranes.[1] This mechanism is a cornerstone of recent theoretical developments based upon Stone's Tensor Surface Harmonic (TSH) theory which have provided powerful general orbital symmetry selection rules for such processes.[2-5] These theories

[1] Lipscomb, W. N. *Science* **1966**, *153*, 373.

[2] Wales, D. J.; Stone, A. J. *Inorg. Chem.* **1987**, *26*, 3845.

[3] Wales, D. J.; Mingos, D. M. P.; Lin, Z. *Inorg. Chem.* **1989**, *28*, 2754.

0002-7863/92/1514-5399$03.00/0 © 1992 American Chemical Society