Distributions in Protein Conformation Space: Implications for Structure Prediction and Entropy

David C. Sullivan and Irwin D. Kuntz
Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-2240

ABSTRACT By considering how polymer structures are distributed in conformation space, we show that it is possible to quantify the difficulty of structural prediction and to provide a measure of progress for prediction calculations. The critical issue is the probability that a conformation is found within a specified distance of another conformer. We address this question by constructing a cumulative distribution function (CDF) for the average probability from observations about its limiting behavior at small displacements and numerical simulations of polyalanine chains. We can use the CDF to estimate the likelihood that a structure prediction is better than random chance. For example, the chance of randomly predicting the native backbone structure of a 150-amino-acid protein to low resolution, say within 6 Å, is 10^{-14}. A high-resolution structural prediction, say to 2 Å, is immensely more difficult (10^{-57}). With additional assumptions, the CDF yields the conformational entropy of protein folding from native-state coordinate variance. Or, using values of the conformational entropy change on folding, we can estimate the native state's conformational span. For example, for a 150-mer protein, equilibrium α-carbon displacements in the native ensemble would be 0.3–0.5 Å based on $T\Delta S$ of 1.42 kcal/(mol residue).

INTRODUCTION

Macromolecules have a large number of internal degrees of freedom. Their conformation space is of high dimension and unevenly populated. The six rigid degrees of freedom (translation and rotation about the center of mass), which figure prominently in the configurational theories of simple fluids, are relatively unimportant. In earlier work, we have shown that simple representations of the distribution of geometric differences among conformers can be used to analyze experimental data and models of protein structure, protein-folding kinetics, and, most recently, the information content of lattice models of proteins (Sullivan and Kuntz, 2001, 2002; Sullivan et al., 2003). In this article we focus on construction of a numerical form for this distribution function for models of protein chains. We can use the distribution function to assess the probability of a conformer lying within a given distance of another conformer and the number of ''effective'' degrees of freedom that operate at that distance. With some standard assumptions, we can estimate the change in conformational entropy upon protein folding.

First consider structure prediction. Previous efforts at assessing the significance of a prediction of native protein conformations use the distribution of conformational differences among a set of random conformations to provide a comparison with a conformation representing the native state (Cohen and Sternberg, 1980; Reva et al., 1998; Feldman and Hogue, 2002). From the differences among random conformations one constructs a cumulative distribution function (CDF). This function presents the probability of two randomly selected conformations being separated by less than a conformational distance, $P(R < r)$. A common measure of conformational distance is the RMS distance between corresponding α-carbon positions after optimal superposition (Levitt, 1976), which we use here and refer to simply as the root-mean-square deviation (RMSD). Specifically, we construct the CDF from a conformational ensemble of W members as:

$$v(r) = \frac{1}{W}\sum_{i=1}^{W}\frac{1}{W-1}\sum_{j=1}^{W}\begin{cases}1 & C\alpha - RMSD^{i,j} \le r \text{ and } i \ne j\\ 0 & \text{otherwise}\end{cases}. \quad (1)$$

The CDF (or its differentiated form, the probability density function) can be extrapolated to the very rare events at low RMSD by fitting the observed distribution to a model of conformer distribution. This task might seem straightforward. However, in the absence of analytical models, finding probabilities for low RMSD conformations requires extensive extrapolation of the CDF. For a medium-sized protein, direct numerical sampling can only access probabilities at the resolution of protein folds (>6 Å) using present computing power. Although the probability of randomly generating a near-native conformation is extremely low (see below), this probability provides a critical measure of the information imparted by prediction schemes.

It is of interest to know how much a particular prediction scheme constrains the set of acceptable conformations. Information gained in structure prediction depends on what is known a priori about a particular protein (i.e., sequence-derived fold family, experimental constraints, ''hard'' force-field constraints on bond lengths and angles). To examine

Submitted February 18, 2004, and accepted for publication March 23, 2004.
Address reprint requests to Irwin D. Kuntz, Dept. of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-2240. Tel.: 415-476-1937; Fax: 415-502-1411; E-mail: kuntz@cgl.ucsf.edu.

0006-3495/04/07/113/08 $2.00
doi: 10.1529/biophysj.104.041723

these prior constraints, we initially consider very basic chain models, though our methods can readily be applied to more realistic constraint sets. The first ensemble we treat is a polypeptide with random ϕ/ψ-dihedral angles. The random polypeptide set (RPP) has fixed bond lengths and bond angles whereas the backbone torsion angles ϕ and ψ are random and unbiased. We choose such a simple system to help understand the underlying form of the distribution functions and provide a basic reference state for measuring the effects of additional structural constraints. Our approach is analogous to developing thermodynamic relationships from an ideal gas model before extending to van der Waals gases or real gases. Enforcing compactness and excluded chain volume lead to more complex CDF functions, as shown below.

Assuming conformational ensembles behave as point sets drawn from a continuous space whose dimensionality is equal to the number of degrees of freedom imposes limits to the functional form of the CDF. We know that CDFs depend very strongly on distance at very small distances. That is, the limiting slope of probability as a function of conformational distance is exponentially steep. From simple geometric arguments (see Sullivan and Kuntz, 2001; Stark et al., 2003; and below), the limiting slope on a double logarithmic plot of a CDF approaches the number of mechanical degrees of freedom at small displacement. In this work, we also find that the log-log slope, appropriately normalized for the number of degrees of freedom and polypeptide volume, is surprisingly independent of chain length for a given constraint set at small RMSD. We validate this behavior for small chain lengths where it is possible to sample conformational space thoroughly. We then use this limiting behavior to approximate the limiting log-log slope functions of longer chain ensembles in the low-RMSD regime. We next extend the CDF functions using Euler extrapolation.

In a final step, we ask whether these CDF extrapolations match standard distribution functions that have been proposed in earlier work. Specifically, the extreme-value distribution (*EVD*) and integrated normal error distribution (*INED*) have been used in previous studies (Cohen and Sternberg, 1980; Reva et al., 1998; Feldman and Hogue, 2002). Levitt and Gerstein (1998) have proposed distribution functions for a similar problem. As a final step in the consideration of appropriate probability distribution functions, we examine the integrated radial density function of a random walk in a high dimensional space (*IRW*). This formulation is motivated by Flory's (1953) treatment of the distribution of end-to-end distances in polymer chains. Our extension to conformational distributions requires one Gaussian for each conformational degree of freedom with the width of the Gaussian representing the displacement magnitude associated with that degree of freedom.

A closely related topic is the calculation of entropy from the variance in atomic coordinates. An exact formula exists for small displacements if the variance is assumed to arise from harmonic vibrations (Levy et al., 1984). The procedure for ensembles that include geometric variation due to conformational events is more challenging. Native-state atomic variances can be modeled by molecular dynamics and are available, in principle, from NMR (Philippopoulos and Lim, 1999) and crystallographic data. Similarly, the unfolded state can be approximated as a random coil, or as a random coil constrained by experimental data for a particular protein (Choy and Forman-Kay, 2001). However, calculating an entropy difference between two states by a simple combination of individual atomic variances is not useful because the atom displacements are strongly intercorrelated. Here we consider the special case where the set of conformations from one thermodynamic state (the reference state) is used to construct the CDF. If a second state can be defined as a subset of the reference state and has all its conformations within some radius of an arbitrary conformation then the statistical entropy difference between the two states is $RT\ln(\mathrm{CDF}(r))$. If the reference state of random conformations includes both the unfolded state and the native state, and if the native state is all conformations within a particular known displacement radius about a single conformation, then this relationship can be used to estimate the conformational entropy of unfolding.

THEORY AND METHODS

Ensemble generation

Three different ensemble types (constraint sets) are analyzed here. 1), The random ϕ/ψ polypeptide ensembles use canonical amino acid bond lengths and bond angles, with the ϕ- and ψ-torsion angles uniformly random over 0–360°. The ω-torsion angle is fixed at 180°. These ensembles do not include excluded volume constraints. We add excluded volume by generating polyalanine conformations using the YARN program (Gregoret and Cohen, 1991). We develop two interesting polyalanine ensembles: 2), the extended ensembles (YARN-EX) have no compactness constraint imposed; and 3), compact ensembles (YARN-C) are constrained during generation to fit inside an ellipsoid boundary with volume $123.9N$ Å^3. Unless stated otherwise, ensembles have a size (W) of 10,000 members for $N = 3$, 4, 5, 6, 30, 50, 70, 100, and 150, and 30,000 members for $N = 8$, 10, 12, 15, 20, and 25, where N is the number of residues per chain. We use higher sampling for medium chain lengths to compensate for added degrees of freedom relative to short chain lengths. Increased sampling for longer chain lengths was computationally too expensive.

Distribution functions

Calculation of v(r)

For a given ensemble, the observed cumulative distribution, $v(r)$, is constructed by calculating the Cα-RMSD (Martin, 1998) for all conformational pairs and integrating the results over the W conformers in the set (Eq. 1). Although the RMSD, calculated in this way, is known not to be a proper metric (Crippen and Ohkubo, 1998), it is sufficient for our purposes (Sullivan and Kuntz, 2001). Our extrapolation of $v(r)$ will use the value of its logarithmic slope, $n(r)$. We define $n(r)$ as $d(\log(v(r)) / d(\log(r))$ and calculate it using the slope from a least-squares line fit over a neighborhood of $v(r)$ points about r, with $v(r)$ and r first transformed to a logarithmic scale. The small-r tails of $v(r)$ are generally noisy; the $n(r)$ plots are calculated from $v(r)$

with the lowest two orders of magnitude (two log units) truncated. For example, for a 10,000 member ensemble, $v(r)$ will generally be defined down to $v(r) = 2 \times 10^{-8}$. However, only the $v(r)$ segment $>2 \times 10^{-6}$ will be used for calculating $n(r)$.

The slope, $n(r)$, has physical significance because it is equal to the number of degrees of freedom of the ensemble in the limit of small RMSD (Sullivan and Kuntz, 2001). This relationship can be generated analytically and empirically for two-dimensional lattice walks (Sullivan et al., 2003). It is also supported by numerical experiments on three-dimensional polymer chains in this article and is consistent with a simple geometric interpretation. Consider the volume behavior (vol(r)) of a hyperdimensional sphere:

$$\mathrm{vol}(r) = Cr^{\mathrm{dim}}, \tag{2}$$

where r is the radius, *dim* the dimensionality of the sphere, and C the hyperdimensional content of a unit-radius hypersphere. Equating the logarithms of both sides of Eq. 2:

$$\log[\mathrm{vol}(r)] = \log(C) + dim\log(r). \tag{3}$$

In a plot of log(r) vs. log(vol), the slope equals the dimensionality of the sphere. By extension, the slope of log($v(r)$) vs. log(r) [$n(r)$] is related to the dimensionality of conformational space, that is, the number of mechanical degrees of freedom. The relationship is not rigorous because the populated region of conformational space is bounded and hence does not necessarily grow in such a simple manner. However, we report, below, that, in the limit of small displacements, conformational space does increase about an arbitrary conformation with a growth exponent, $n(r)$, equal to the number of degrees of freedom.

An intuitive understanding of $n(r)$, or *dim* in Eq. 2, can be gained by considering a conformational space shaped as a long cylindrical rod (three dimensions) (Sullivan et al., 2003). The r-dependent volume about a reference point embedded inside the rod only increases as $\sim r^3$ at radii less than the cylinder's radius. At longer radial scales the rod behaves as a one-dimensional object with the volume function approaching $\sim r^1$. The volume of rod bound by radii longer than the rod's length is constant with no radius dependence, $\sim r^0$. This example illustrates how $n(r)$ reports the number of degrees of freedom (dimensionality) effective at a resolution of r.

Assuming $v(r)$ is known to some small RMSD value, and additionally that $n(r)$ is known (or can be estimated) near $r = 0$ (see below), we can perform an Euler extrapolation of $v(r)$ to any nearby value of RMSD. In practice, $v(r)$ is first transformed to a logarithmic scale and then iteratively linearly extrapolated to smaller r over a small increment ($\Delta\log(r) = 0.0001$) using a slope of $n(r)$. This provides a new log(v(log(r))) point for extrapolation with an updated value for n. We refer to the Euler extrapolated CDF as $v_e(r)$.

Extreme-value distribution

The extreme-value distribution (*EVD*) used by Feldman and Hogue (2002) has the form

$$EVD(r) = \exp(-\exp((x - r)/w)). \tag{4}$$

Equation 4 defines one of three types of extreme-value distributions, referred to as a Gumbel type-I distribution. We derived the parameters by least-squares fitting ln($v(r)$) to $-\exp((x - r) / w)$, where x and w are fitting parameters.

Integrated normal error distribution

The integrated normal error distribution (*INED*) was calculated using the mean (μ) and variance (σ^2) of the RMSD distributions and numerically integrating the normal error distribution function:

$$P(r) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-0.5\left(\frac{r-\mu}{\sigma}\right)^2\right). \tag{5}$$

Integration of Eq. 5, which tails off to negative and positive infinity, was initiated at $r = -10$ Å.

Levitt and Gerstein distribution

Levitt and Gerstein (1998) describe a probability density function that essentially is an inverted parabola on a log-log scale. The Levitt and Gerstein density function (*LGD*) is

$$LGD(r) = \exp\left(-\left(\frac{\ln r - C_1}{C_2}\right)^4\right). \tag{6}$$

The fitting parameters, C_1 and C_2, are found by fitting the natural logarithm of the observed density function to ln(*LGD*). The corresponding CDF, *ILGD*, is found by numerically integrating *LGD*. We also examine a similar function using a square-power in the exponent, which we term *LGD*2

$$LGD2(r) = \exp\left(-\left(\frac{\ln r - C_1}{C_2}\right)^2\right). \tag{7}$$

We plot the integrated form, *ILGD*2, by numerical integration of *LGD*2 using best-fit parameters for C_1 and C_2.

D-dimensional random walk displacement distribution

The D-dimensional random walk displacement cumulative distribution (*IRW*) is found by integrating its respective radial probability density function. The three-dimensional radial probability density function, which is used by Flory (1953) to model the distribution of distances between two beads of a freely jointed chain, is given by:

$$g_{D=3}(r) = \left(\frac{\sqrt{3}}{\sqrt{\langle R^2\rangle}\sqrt{2\pi}}\right)^3 \exp\left(-\left(\frac{3r^2}{2\langle R^2\rangle}\right)\right)4\pi r^2. \tag{8}$$

This function is arrived at by first taking the product of three normal distributions, yielding a spatial probability density function, and multiplying by a sphere's surface area, $4\pi r^2$, to give the radial probability density function. $\langle R^2\rangle$ is the mean-squared displacement for the walk. By analogy, the D-dimensional radial probability density function is:

$$g_D(r) = \left(\frac{\sqrt{D}}{\sqrt{\langle R^2\rangle}\sqrt{2\pi}}\right)^D \exp\left(-\left(\frac{Dr^2}{2\langle R^2\rangle}\right)\right)DV_D r^{D-1}. \tag{9}$$

The $DV_D r^{D-1}$ term is the surface area of a D-dimensional sphere of radius r where V_D equals the volume of a unit-radius D-dimensional sphere (Conway and Sloane, 1988):

$$V_D = \frac{\pi^2}{2\int_0^\infty \exp(-x^2)x^{D+1}dx}. \tag{10}$$

V_D, then *IRW*(r), are found by numerical integration.

RESULTS

The CDF log-log slope, $n(r)$

For polyalanine ensembles generated using the constraint sets considered here (RPP, YARN-EX, and YARN-C), $n(r)$ convincingly converges on $2N - 5$ as r approaches 0, provided that N is small enough for the sampling to be sufficient (Fig. 1). Longer chains would require exponentially more sampling to define $n(r)$ near $r = 0$. However, the data we have indicate that all the curves trend toward $2N - 5$ at $r = 0$.

The primary difference between the ensembles that obey excluded volume constraints (YARN-EX, YARN-C) and the ensembles that do not obey excluded volume (RPP), is a dip in $n(r)$ at ~1–2 Å in the curves for the former sets with the dip being more pronounced in the YARN-C data. The most likely source of this dip is that the excluded volume constraints cause depletion in conformations that are within a radius about accepted conformations (I. Kuntz and D. Sullivan, unpublished data). This feature is thus analogous to the characteristic packing defects in normalized radial distribution functions of liquids.

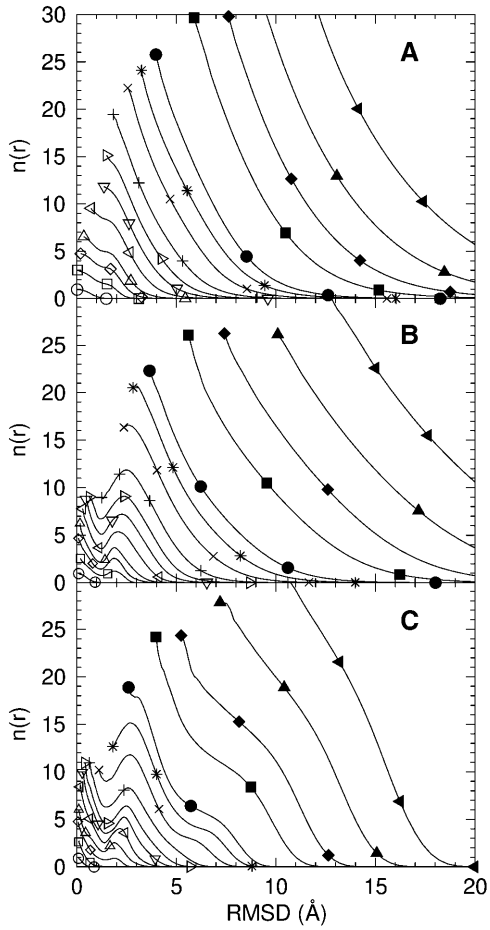

FIGURE 1 $n(r)$ computed for $N = 3$ (○), 4 (□), 5 (◇), 6 (△), 8 (◁), 10 (▽), 12 (▷), 15 (+), 20 (×), 25 (★), 30 (●), 50 (■), 70 (♦), 100 (▲), and 150 (◀) for (*A*) RPP, (*B*) YARN-EX, and (*C*) YARN-C ensembles.

The compactness constraint in YARN-C also reduces the range of possible conformational displacements, which shifts the $n(r)$ curves to smaller r. Additionally, there is a steeper descent of $n(r)$ from its limiting value of $2N - 5$ at $r = 0$ relative to the YARN-EX curves.

Extracting a "universal curve" for $n(r)$

We seek simple scaling functions that reduce the dependence of $n(r)$ on chain length. The first step is to normalize for the number of degrees of freedom:

$$n_{\text{norm}}(r) = n(r)/(2N - 5). \tag{11}$$

This normalization does not remove all differences across ensembles (data not shown). Curves for the small-N ensembles deviate from the larger-N $n_{\text{norm}}(r)$ curves, particularly for YARN-EX and YARN-C. The deviation may reflect a qualitative geometric feature of small chain-length conformations, such as the lack of complete globularity or being relatively more constrained by covalent bonds than nonlocal interactions. Ignoring the curves with $N < 8$, the $n_{\text{norm}}(r)$ curves superimpose well up to ~2 Å. At larger RMSDs, where we see additional dependence on chain length, the polypeptide volume dependence must be considered (Maiorov and Crippen, 1994). In Fig. 2, the RMSD values >~2 Å have been divided by $\sim N^{1/3}$. Specifically, the scaled RMSD (sRMSD) is:

$$sRMSD = \begin{cases} RMSD & RMSD \le p_1 \\ p_1 + ((RMSD - p_1)(N - p_2)^{p_3}) & RMSD > p_1 \end{cases}. \tag{12}$$

This scaling equation, defined only for larger N ($N > p_2$), was parameterized by examining RMSD N-dependence at fixed $n_{\text{norm}}(r)$ values. Fitting parameters are listed in Table 1. Using this normalization, the $n_{\text{norm}}(r)$ curves superimpose well except for a small spread at the largest sRMSD values.

We use these scaled curves (Eq. 12) to extrapolate the $n_{\text{norm}}(r)$ curves to low sRMSD by iterative concatenation of the $n_{\text{norm}}(r)$ segment corresponding to the next smallest N. For example, the extrapolation of $n_{\text{norm}}(r)$ for YARN-C, $N = 70$, which is defined only to sRMSD = 2.62, is initiated by appending the low sRMSD segment of $N = 50$'s $n_{\text{norm}}(r)$, thus defining the extrapolation to sRMSD = 2.38. This is repeated down to the terminus of the $N = 8$ ensemble's $n_{\text{norm}}(r)$ curve at sRMSD = 0.16, $n_{\text{norm}}(r) = 0.77$. A linear extrapolation to zero RMSD, unity $n_{\text{norm}}(r)$, is used below this value. These extrapolated $n_{\text{norm}}(r)$ curves are converted back to $n(r)$ by the inverse function of Eq. 12 and are then used to construct their respective $v_e(r)$ functions (Fig. 3). Table 2 provides a tabulated form of the logarithm of $v_e(r)$. For any listed threshold, r, in Table 2 (i.e., any column),

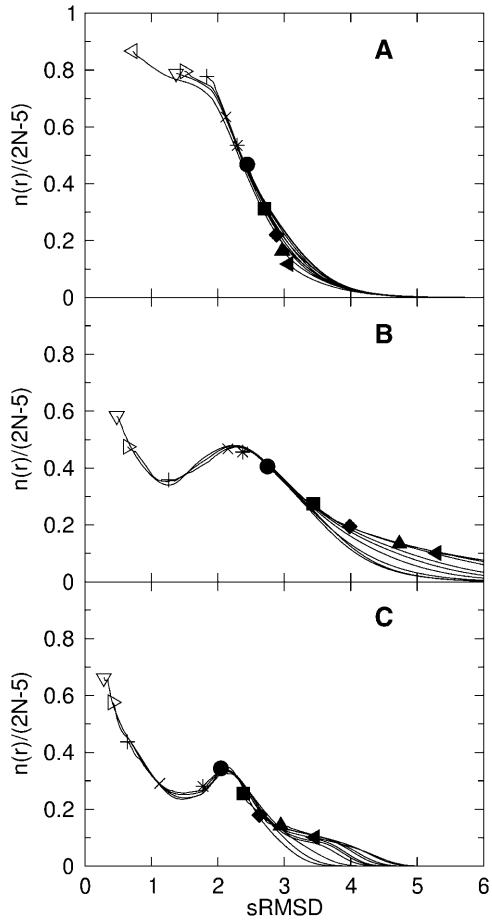

FIGURE 2 $n_{norm}(r)$ for (*A*) RPP, (*B*) YARN-EX, and (*C*) YARN-C is plotted as a function of sRMSD, as defined in Eq. 12 using the appropriate constraint set's fitting parameters listed in Table 1. Chain length is indicated using Fig. 1's symbol mapping. For clarity, only the initial point of each curve has a symbol designation. The smallest-*N* curves are not shown.

correlation (r^2) of $-\log_{10}(v_e(r))$ with N is >0.988, using data to full precision, suggesting that the full range of N can be safely linearly interpolated.

The simplest interpretation of these results is that the chances of random generation of even low-resolution structures of proteins ($N > 100$) are very small and the random generation of good quality structures is out of the question with current computational resources. On the other hand, the distributions depend so steeply on the chain length, that Table 2 suggests that random exploration of small chains ($N \sim 30$ residues) might provide positive results if a suitable scoring function is available (Feldman and Hogue, 2002). Of course, most prediction schema depend on nonrandom constraints. Our approach provides a way to compare different models by comparing the limiting behavior of their respective CDFs and hence their number of effective degrees of freedom.

TABLE 1 Fitting parameters for Eq. 12

Ensemble	p_1	p_2	p_3
RPP	1.91	5.4	−0.43
YARN-EX	1.87	8.6	−0.23
YARN-C	1.75	5.6	−0.33

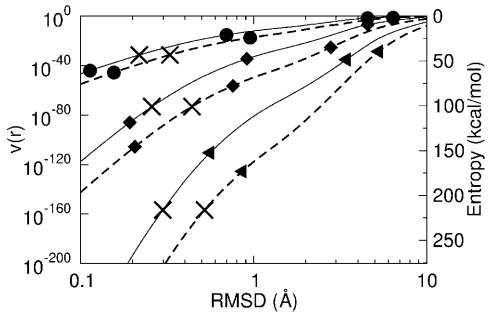

FIGURE 3 $v_e(r)$ for YARN-C (*solid lines*) and YARN-EX (*dashed lines*) is plotted for $N = 30$ (●), 70 (♦), and 150 (◀). $v_e(r)$ is constructed from the observed distributions, $v(r)$, which terminate at 10^{-6}. Entropy equivalency ($T\Delta S_{conf}$) for $v(r)$ is listed on the right-hand axis for $T = 298$ K in units of kcal/mol. The point where $v_e(r)$'s derived entropy equals 1.42 kcal/(mol residue) (see text) is marked by ''×''.

Conformational entropy of protein folding

The cumulative distribution function provides the relative probability of finding a conformation within a particular distance of another conformation, averaged over the entire ensemble. It applies to ensembles made of discrete conformers (e.g., lattice models or off-lattice chains whose geometries are at minima on an energy surface) and to continuously variable geometries such as the YARN models for which no energy separations are used. For thermodynamic states that are geometrically ''nested,'' the CDF provides a direct means for calculating the conformational entropy between states:

$$T\Delta S_{conf} = -RT\ln(v(r)). \quad (13)$$

The right-hand axis in Fig. 3 gives this quantity in kcal/mol for $T = 298$ K. Equation 13 might be useful for characterizing the entropy change on protein unfolding if the native state could be defined as all conformations within some radius, r_{native}, of a particular conformation (i.e., the global minimum) and to reside within the constraint boundaries of a larger ''unfolded'' conformational space with a characterized $v(r)$. Under these assumptions, the conformational entropy upon protein folding, $T\Delta S_{conf,fol}$, is simply $-RT\ln(v(r_{native}))$. There are several methods for estimating the magnitude of native-state thermal displacement, r_{native}. The backbone RMSD between high-resolution crystal structures of identical proteins from different crystal environments, generally ~0.4 Å (Chothia and Lesk, 1986),

TABLE 2 Expectation values [$-\log_{10}(v_e(r))$] for extended (YARN-EX) and compact (YARN-C) ensembles

	YARN-EX						YARN-C					
N	0.2 Å	0.5 Å	1 Å	2 Å	3 Å	6 Å	0.2 Å	0.5 Å	1 Å	2 Å	3 Å	6 Å
30	40	25	17	11	7	1.5	33	19	12	7.3	4.2	0.7
70	107	69	50	34	23	7.7	84	50	33	22	14	4.1
100	157	102	74	52	36	12	125	75	51	35	24	7.7
150	243	160	118	84	61	23	194	117	81	57	41	14

provides one estimate. Displacement among a conformational ensemble modeled using NMR data, generally >1 Å, provides another estimate based on more relevant solution state data, although generally these models are based on less experimental data and more ad hoc mathematical assumptions than x-ray structures. Molecular dynamics trajectories on the native state provide a third means for estimating r_{native}, with variances >1 Å, if one accepts inherent limitations in the force field and sampling (Troyer and Cohen, 1995). Neglecting for the moment debate over quality differences between x-ray and NMR models (Lee and Kollman, 2001), the associated entropy difference between $r_{native} = 0.4$ Å and 1.0 Å is ~80 kcal/mol for the 150-mer (Fig. 3), or ~0.5 kcal/mol/residue at 298 K.

Inversely, r_{native} can be arrived at from estimates of $\Delta S_{conf,fol}$. A recent literature survey, with an accompanying experimental measure, place this value at ~.00478 (kcal)/(mol K residue) (Thompson et al., 2002), equal to 1.42 (kcal)/(mol residue) at 298 K. This value, multiplied by N, is indicated on the $v_e(r)$ curves for $N = 30$ ($r_{native} = 0.22$–0.33 Å, 43 kcal/mol), 70 ($r_{native} = 0.26$–0.44 Å, 100 kcal/mol), and 150 ($r_{native} = 0.30$–0.52 Å, 213 kcal/mol). The high r_{native} value comes from YARN-EX and the low r_{native} value comes from YARN-C data. If YARN-EX is assumed to overrepresent the variance of the unfolded state and if YARN-C provides an over-constrained model of the unfolded state, then the true r_{native} should lie within our given RMSD range. This calculation has the caveat that the actual CDF function for any particular protein's unfolded state will be sequence dependent. Even if the sequence is specified, models for unfolded states of proteins, in general, are a point of considerable debate in the literature (Baldwin and Zimm, 2000; Choy and Forman-Kay, 2001; Plaxco and Gross, 2001; Shortle and Ackerman, 2001; van Gunsteren et al., 2001; Goldenberg, 2003). Our results can only capture general behavior and provide guiding principles.

Comparison of models for the CDF function

Others (Cohen and Sternberg, 1980; Levitt and Gerstein, 1998; Reva et al., 1998; Feldman and Hogue, 2002) have modeled the CDF using standard distribution functions. Fig. 4 compares four of these functions (*EVD*, *INED*, *ILGD*, and *ILGD2*) with $v_e(r)$ for the YARN-C, $N = 100$ ensemble. The *INED* very quickly diverges from $v_e(r)$ at RMSD ~ 3 Å. The *EVD* is superior by tracking $v_e(r)$ for orders of magnitude into the extrapolation regime but it fails in a fashion similar to the *INED* by overestimating low-RMSD probabilities and qualitatively fails at low RMSD (~0.5 Å). For the extended conformational ensembles, the *EVD* divergence is at larger RMSD, e.g., for YARN-EX, $N = 100$, *EVD* divergence is ~4 Å, $v_e(r) = 10^{-25}$, and *INED* diverges at ~12 Å, $v(r) = 0.01$. In contrast to *INED* and *EVD*, the *ILGD* and *ILGD2* diverge by becoming too steep, greatly underestimating $v_e(r)$ at small r.

The high-dimensional random walk model leads to a distribution (*IRW*) that is interesting because its form can be related to earlier approaches to coordinate distributions. Dimensionality enters as an explicit parameter. The $\langle R^2 \rangle$ term bears a simple relationship to the variance in each dimension as $\langle R^2 \rangle / D$. Density in each dimension is assumed to be normally distributed. Fig. 5 shows several *IRW* curves, with various values for $\langle R^2 \rangle$ and D, compared to the 150-mer RPP ensemble's $v_e(r)$. The (root-) mean-squared RMSD for this ensemble is 473.2 Å^2 (21.8 Å). Two curves use the proper D of 295 (i.e., $2N - 5$). One of these curves models large variance ($\langle R^2 \rangle^{1/2} = 21.8$ Å) and the other curve models small variance ($\langle R^2 \rangle^{1/2} = 8.5$ Å). Although neither curve accurately fits $v_e(r)$ over the entire given range, it is also true that neither diverges from $v_e(r)$ at small r, in contrast to the other functions explored in Fig. 4. Because both $v_e(r)$ for $N = 150$ and *IRW* for $D = 295$ have the same small-r limiting slope,

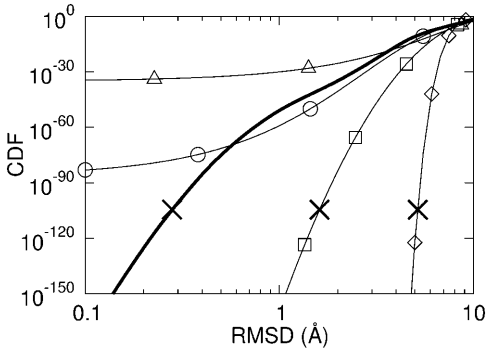

FIGURE 4 YARN-C derived CDF functions $v_e(r)$ (*solid line*, *no symbol*), *EVD* (○), *INED* (△), *ILGD* (◇), and *ILGD2* (□) are plotted for $N = 100$. Function intersections with the conformational entropy of unfolding (Thompson et al., 2002) for a 100-mer protein (equal to 142 kcal/mol, calculated from 1.42 kcal/(mol residue) × 100 residues) are marked by "×".

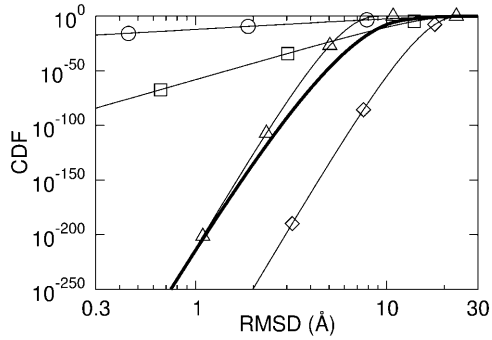

FIGURE 5 Random-walk-based cumulative distribution function (*IRW*) with parameters $\langle R^2\rangle^{1/2} = 21.75$, $D = 10$ (○), $\langle R^2\rangle^{1/2} = 21.75$, $D = 50$ (□), $\langle R^2\rangle^{1/2} = 21.75$, $D = 295$ (◇), $\langle R^2\rangle^{1/2} = 8.5$, $D = 295$ (△). In bold line (no symbol) is $v_e(r)$ for RPP, $N = 150$.

there exists a $\langle R^2\rangle^{1/2}$ value (~8.5 Å) for which *IRW* and $v_e(r)$ converge at small r. In contrast, the large-r portion of $v_e(r)$ is better represented by small dimensionality (~10), large displacement *IRW* curves. The fact that no single *IRW* curve accurately models $v_e(r)$ is not surprising given that the amplitudes of orthogonal displacement modes across a conformational ensemble generally fill a spectrum (Garcia, 1992; Sullivan and Kuntz, 2001).

DISCUSSION

The cumulative distribution function appears to be a useful starting point for exploring the conformational space of polymer chains. It directly addresses the question of the likelihood of a given structure falling within an arbitrary distance of another structure, and, under certain circumstances, it provides a direct route to estimating conformational entropies (see also Sullivan et al., 2003). The distribution functions can be constructed by stochastic sampling or by full enumeration for a variety of polymer models. It is important to know if one of the standard distribution functions is capable of representing the full range of interest for the CDF, which can span >100 orders of magnitude for representations of small proteins. Our initial survey suggests that the answer is no. A multidimensional Gaussian model offers the most promising general approach, but much work needs to be done to understand how to parameterize such a model.

Another interesting point is the regularity in the limiting (log-log) slope of the CDF versus RMSD; this slope is directly related to the number of mechanical degrees of freedom. We make use of this relationship to derive empirical CDF curves for a variety of systems and suggest that it be computed for predictive schemes as a point of comparison. This work underscores the importance of considering the underlying physics, namely, the number of degrees of freedom and the displacement distribution along each dimension.

In our earlier work (Sullivan and Kuntz, 2001) we explored describing ''conformational volume'' with only two parameters that are analogous to D and $\langle R^2\rangle$. In that work (Sullivan and Kuntz, 2001) we examined the probability distribution exclusively on a linear scale (i.e., high probability) and thus found small D, large $\langle R^2\rangle$, best represented conformational space. We later (Sullivan and Kuntz, 2002) tried to apply this result to a simple dynamic model for protein folding as biased diffusion in a high-dimensional box. For that problem we found that a large number of dimensions that permit only small displacement, in addition to the few large displacement dimensions, are required to capture the proper time-displacement behavior of protein dynamics over the unfolded state from femtosecond to microsecond timescales (Sullivan and Kuntz, 2002). In other words, that work optimistically suggests that as few as four parameters can describe conformational distributions, e.g., 1), D; 2), $\langle R^2\rangle_{large}$; 3), $\langle R^2\rangle_{small}$; and 4), fraction of large displacement versus small displacement dimensions. Perhaps a clever combining of random walk curves could similarly be used to model $v_e(r)$ for the RPP ensemble over its entirety. Excluded volume would likely enter into this formalism by subtracting an ''excluded'' distribution function from the parent RPP distribution.

The $v_e(r)$ distribution points to a method of relating entropy changes to structural variance for certain types of conformational constraint sets. Structural models with variances larger than the known experimental entropy must admit to error or offer an additional explanation. For example, describing the native state as a collection of substates distributed in a wider, anisotropic energy basin (Frauenfelder et al., 1991) could explain larger native state displacements than Fig. 3 directly predicts. The constraint sets used here likely only capture a small portion of this description. We can, for instance, modify our native-state model to a collection of many disjoint substates of slightly smaller r, which, in sum, retain the same probability as a single larger one. For example, for the 150-mer extended ensemble (Fig. 3), 1000 substates of $r = 0.50$ Å, or 10^6 substates of $r = 0.48$ Å, have the same probability as r_{native} of 0.52 Å.

In future work, we could include protein-like amino acid sequences and energy relaxation to strengthen connections to real proteins as well as explore side-chain packing effects on main-chain entropy. We are also curious about the effects of energy minimization, which discretizes conformational space, thus introducing a lower limit to the CDF in the vicinity of $v(r) = (\text{total number of minima})^{-1}$, at a resolution of ~0.1 Å (Troyer and Cohen, 1995). Sampling issues limit direct numerical observation of this limit to very short sequences, however this feature could be extrapolated to longer sequences by the methods outlined here. Such a study would help position all-atom molecular models within the context of statistical mechanics, heretofore reserved for simplified lattice representations of proteins (Chan and Dill, 1989).

We are grateful for helpful discussions with Ken Dill and Gordon Crippen. Jerome Nilmeier provided assistance in preliminary one-dimensional simulations.

This work was supported by the National Science Foundation (R. Kip Guy, principal investigator).

REFERENCES

Baldwin, R. L., and R. H. Zimm. 2000. Are denatured proteins ever random coils? *Proc. Natl. Acad. Sci. USA.* 97:12391–12392.

Chan, H. S., and K. A. Dill. 1989. Compact polymers. *Macromolecules.* 22:4559–4573.

Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.

Choy, W. Y., and J. D. Forman-Kay. 2001. Calculation of an ensemble of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* 308:1011–1032.

Cohen, F. E., and J. E. Sternberg. 1980. On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.* 138:321–333.

Conway, J. H., and N. J. A. Sloane. 1988. Sphere packings, lattices and groups. Springer-Verlag, New York.

Crippen, G. M., and Y. Z. Ohkubo. 1998. Statistical mechanics of protein folding by exhaustive enumeration. *Proteins.* 32:425–437.

Feldman, H. J., and C. W. V. Hogue. 2002. Probabilistic sampling of protein conformations: new hope for brute force? *Proteins.* 46:8–23.

Flory, P. J. 1953. Principles of Polymer Chemistry. Cornell University Press, Ithaca, NY.

Frauenfelder, H., S. G. Sligar, and P. G. Wolynes. 1991. The energy landscapes and motions of proteins. *Science.* 254:1598–1603.

Garcia, A. E. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68:2696–2699.

Goldenberg, D. P. 2003. Computational simulation of the statistical properties of unfolded proteins. *J. Mol. Biol.* 326:1615–1633.

Gregoret, L. M., and F. E. Cohen. 1991. Protein folding. Effect of packing density on chain conformation. *J. Mol. Biol.* 219:109–122.

Lee, M. R., and P. A. Kollman. 2001. Free-energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure.* 9:905–916.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.

Levitt, M., and M. Gerstein. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA.* 95:5913–5920.

Levy, R. M., M. Karplus, J. Kushick, and D. Perahia. 1984. Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an alpha-helix. *Macromolecules.* 17:1370–1374.

Maiorov, V. N., and G. M. Crippen. 1994. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* 37:625–634.

Martin, A. C. R. 1998. ProFit 1.8. University College, London, UK.

Philippopoulos, M., and C. Lim. 1999. Exploring the dynamic information content of protein NMR structures: comparison of a MD simulation with the NMR and x-ray structures of *E. Coli* RNase HI. *Proteins.* 36:87–110.

Plaxco, K. W., and M. Gross. 2001. Unfolded yes but random never. *Nat. Struct. Biol.* 8:659–660.

Reva, B. A., A. V. Finkelstein, and J. Skolnick. 1998. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold. Des.* 3:141–147.

Shortle, D., and M. S. Ackerman. 2001. Persistence of native-like topology in a denatured protein in 8 M urea. *Science.* 293:487–489.

Stark, A., S. Sunyaev, and R. B. Russell. 2003. A model for statistical significance of local similarities in structure. *J. Mol. Biol.* 326:1307–1316.

Sullivan, D. C., and I. D. Kuntz. 2001. Conformation spaces of proteins. *Proteins.* 42:495–511.

Sullivan, D. C., and I. D. Kuntz. 2002. Protein folding as biased conformational diffusion. *J. Phys. Chem. B.* 106:3255–3262.

Sullivan, D. C., T. Aynechi, V. V. Voelz, and I. D. Kuntz. 2003. Information content of molecular structures. *Biophys. J.* 85:174–190.

Thompson, J. B., H. G. Hansma, P. K. Hansma, and K. W. Plaxco. 2002. The backbone conformational entropy of protein folding: experimental measures from atomic force microscopy. *J. Mol. Biol.* 322:645–652.

Troyer, J. M., and F. E. Cohen. 1995. Protein conformational landscapes: energy minimization and clustering of a long molecular dynamics trajectory. *Proteins.* 23:97–110.

van Gunsteren, W. F., R. Burgi, C. Peter, and X. Daura. 2001. The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed.* 40:352–355.