Side-chain entropy and packing in proteins

SARINA BROMBERG AND KEN A. DILL

Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-1204

(RECEIVED January 21, 1994; ACCEPTED April 11, 1994)

Abstract

What role does side-chain packing play in protein stability and structure? To address this question, we compare a lattice model with side chains (SCM) to a linear lattice model without side chains (LCM). Self-avoiding configurations are enumerated in 2 and 3 dimensions exhaustively for short chains and by Monte Carlo sampling for chains up to 50 main-chain monomers long. This comparison shows that (1) side-chain degrees of freedom increase the entropy of open conformations, but side-chain steric exclusion decreases the entropy of compact conformations, thus producing a substantial entropy that opposes folding; (2) there is a side-chain “freezing” or ordering, i.e., a sharp decrease in entropy, near maximum compactness; and (3) the different types of contacts among side chains (*s*) and main-chain elements (*m*) have different frequencies, and the frequencies have different dependencies on compactness. *mm* contacts contribute significantly only at high densities, suggesting that main-chain hydrogen bonding in proteins may be promoted by compactness. The distributions of *mm*, *ms*, and *ss* contacts in compact SCM configurations are similar to the distributions in protein structures in the Brookhaven Protein Data Bank. We propose that packing in proteins is more like the packing of nuts and bolts in a jar than like the pairwise matching of jigsaw puzzle pieces.

Keywords: conformational entropy; lattice model; protein stability; side-chain freezing; side-chain packing

When a polymer becomes compact, as when a protein folds to its native state, the main chain loses degrees of freedom, incurring a large loss of entropy. Main-chain configurational entropy is a strong force opposing the stability of native proteins (Dill, 1985, 1990). How do side chains contribute to the configurational entropy of protein conformations? Is there “side-chain freezing,” i.e., a large sharp loss of degrees of freedom as the side chains pack into compact native conformations? We explore these issues with a simple model that can be studied rigorously.

The inference that side-chain packing is important to protein stability has been drawn from several lines of evidence. Side chains in the hydrophobic cores of proteins are tightly packed (Richards, 1974, 1977; Richards & Lim, 1994) and proteins have low compressibilities (Klapper, 1971; Eden et al., 1982; Gavish et al., 1983; Kundrot & Richards, 1987). Many core side-chain motions are hindered or eliminated in native proteins (Gurd & Rothgeb, 1979; Wagner, 1983; McCammon & Harvey, 1987). Side-chain packing may also be important to the kinetics of folding. Ptitsyn (1987) and Ptitsyn et al. (1990) have suggested that the slow rate of protein folding may be due to the process of packing core side chains. Recent computer simulation and protein design results also suggest that the packing of side chains

is critical. Rey and Skolnick (1993) have shown that the addition of model side chains to a model main chain in a computer simulation of protein folding reduces the number of pathways that lead to the native state, and Handel et al. (1993) have shown that it is not easy to design sequences that can fold to conformations with immobilized side chains.

We can imagine a range of models of how side-chain packing influences protein properties. On the one hand, side-chain packing might resemble lock-and-key fits in which there is specific pairwise matching of complementary side-chain shapes. At the other extreme, side-chain packing might be nonspecific. These views are embodied in 2 extreme models for side-chain packing: a *jigsaw puzzle* model, in which side chains match with specific shape complementarity, and a *nuts and bolts* model, in which side chains pack together without specificity (see Fig. 1). The defining characteristic of the jigsaw puzzle model is a requirement for a *critical point of disjuncture* between the parts. That is, as the protein expands in denaturation, the side chains gain little additional rotational freedom until they reach a particular separation, the critical disjuncture density, where they then suddenly become orientationally free of each other. A quantitative model of this type for side-chain freedom in proteins has been given by Shakhnovich and Finkelstein (1989).

Nuts and bolts packing arises from multiple nonspecific interactions, as in the jumble of nuts and bolts in a jar. This is the way molecules are organized in liquid and glassy states (Ziman,

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143-1204; e-mail: dill@maxwell.ucsf.edu.

Fig. 1. Alternative models for the packing of unconnected side chains. **A:** Jigsaw puzzle model elements fit together with precise complementarity and restrict each other's freedom of motion until they are separated by a critical distance. **B:** Nuts and bolts model elements have no shape complementarity and no critical separation. They gain the freedom to rotate near maximum density.

1979), where there is no requirement for complementarity in the shapes of the objects. Even though the nuts and bolts model lacks pairwise specificity, each piece defines a surrounding "cavity" of its own general size and shape. We define nuts and bolts models to be those for which only a small reduction in the density from the maximally compact state leads to large changes in orientational freedom. In contrast, in jigsaw puzzle models, reducing the density from the maximally packed state at first produces no change in orientational freedom.

In the present study, we do not attempt to model possible pairwise shape complementarity among side chains. Rather, we consider a minimal model of what side chains *must* do, by virtue of the volumes and degrees of freedom they add when they are appended to a main chain. We explore the consequences of these properties through exact enumeration and Monte Carlo (MC) sampling and compare the results with experiments. We find this model falls into the nuts and bolts class.

Our study pertains to a subclass of branched polymers called "regular combs." There have been many calculations of the radii of gyration and solution properties of branched chain polymers with different branching geometries (Casassa & Berry, 1966; Gupta & Forsman, 1973; Solc, 1973; Daoud & Joanny, 1981; Wilkinson, 1986; Nemirovsky et al., 1987; Lipson, 1991) since the pioneering work of Zimm and Stockmayer (1949). Regular combs have identical side groups attached at repeating intervals to a polymer main chain. The radii of comb polymers have been calculated by Solc (1973), McCrackin and Mazur (1981), and Lipson (1991) and measured by Roovers (1979). There are also several studies of the collapse thermodynamics and phase diagrams for some types of branched polymers (Derida & Herrmann, 1983; Dickman & Schieve, 1984; Chang & Shapir, 1988; Lam, 1988; Madras et al., 1990; Gaunt, 1991; Flesia et al., 1992). These works, including the collapse theories,

study chain conformations in or near "theta" or good solvent conditions (Flory, 1953), where conformations are highly expanded. Under such conditions, many properties of polymers are dependent only on long spatial wavelength scales, and the structural details such as side chains are subsumed under the property of chain stiffness.

Thus, polymers including proteins are often modeled as linear strings of beads. In such cases, each bead represents a monomer without further detail (Flory, 1953; Domb, 1963; deGennes, 1979). Can models this simple capture the essential conformational properties of proteins? How do side groups affect conformational properties?

In compact polymers such as native proteins, shorter spatial wavelength properties such as side-chain packing may be important. There are theoretical studies of side-chain entropies in proteins. Creamer and Rose (1992) have performed MC simulations using the AMBER/OPLS force fields and attribute the differences in the helical propensities of amino acids to local helix main-chain interference with side-chain rotations. Our focus here is on compact states, rather than helical structures. Pickett and Sternberg (1993) have developed an empirical scale for the torsional entropies of the 20 different amino acids, omitting the excluded volume entropies of the main chain and the side chains. Here we focus on the effects of excluded volume.

The models

We compare 2 models, shown in Figure 2. The linear chain model (LCM) is our reference (Chan & Dill, 1989a, 1989b; Lau & Dill, 1989). Taking the LCM to represent the main chain, the side-chain model (SCM) is created by attaching a single side-

Fig. 2. Examples of chain configurations in the LCM (top) and SCM (bottom) are shown: ■, main-chain (*m*) monomer; □, side-chain (*s*) monomer. A topological contact is formed when any 2 unconnected monomers occupy adjacent lattice sites. For a chain of length n , main-chain monomers, and $n - 1$ main-chain bonds, the main-chain monomers are numbered $i = 1, 2, 3, \dots, n$. The side chain attached to main-chain monomer i is identified as side-chain monomer i . The figure defines local and nonlocal contacts of all types. Note that topological contacts are local if the contacting monomers i, j are connected by 3 bonds (including main-chain to side-chain bonds for *ms* and *ss* contacts). All other topological contacts are nonlocal.

chain monomer to each main-chain monomer (except the ends; see Methods). This simple representation approximates the frequency-weighted (Karlin et al., 1992) ratio of side-chain to main-chain volumes of the amino acid residues found in proteins (Finney, 1975). Each side-chain monomer has the freedom to occupy any 1 empty lattice site adjacent to its corresponding main-chain monomer. The freedom of a side chain to adopt any orientation represents side-chain rotameric degrees of freedom at a simple level. In real polypeptides, the orientations of the proximal bonds of side chains are determined by the main-chain dihedral angles, but distal side-chain bonds have more freedom. In the SCM, all the side-chain orientational degrees of freedom are represented by the freedom of a single side-chain monomer to occupy any free site next to the main chain. No energetic interactions are incorporated in this study beyond the steric constraint that no 2 monomers (main-chain or side-chain) may occupy the same lattice site. We explore all configurations of the main chain and the side chains that do not violate steric constraints (exhaustively for short chains and by MC sampling for long chains) on the 2-dimensional (2D) square lattice and the 3D simple cubic lattice.

It is useful to study the SCM in 2D as well as 3D because 2D models have been used extensively to represent proteins, due to the computational and physical advantages they offer. The main advantage of 2D models is that they capture the correct perimeter/area ratio for chains sufficiently short that the partition function and other properties can be computed without mathematical approximation. A central question in evaluating the results of model studies is what properties are similar and different in 2D versus 3D. We find that the results of 2D lattice simulations are generally similar to results in 3D, for the SCM as well as the LCM, although we also discuss a few exceptions. Incorporation of specific monomer sequences and sticking energies in the 2D LCM models leads to the following properties of proteins: (1) For a substantial fraction of sequences, small changes in conditions near a point of marginal stability lead to a sharp folding transition, from a large ensemble of denatured conformations to a small number of highly compact conformations with a hydrophobic core (Lau & Dill, 1989, 1990; Chan & Dill, 1991). (2) The compact configurations have length distributions of helices and parallel and antiparallel sheets similar to the protein structures in the Brookhaven Protein Data Bank (PDB) (Chan & Dill, 1990a). (3) The following protein mutational properties are found: (a) The surface monomers are more neutral to mutation than core monomers. (b) There are second-site revertants. (c) A given native structure can be encoded in many different sequences (Lau & Dill, 1990; Chan & Dill, 1991; Shortle et al., 1992). (4) There are favored folding pathways that are sequence-dependent (Miller et al., 1992; Chan & Dill, 1994).

Why use simplified models? What can we learn from them? Higher resolution models invariably require arbitrary parameters and/or limited sampling of conformational space, and sometimes require ad hoc assumptions. For studying the properties of polymers and other materials, it has often been of value to study exact models in order to distinguish the properties that follow directly from physical principles from the properties that are artifacts of mathematical approximations. Our interest here is in the nature of the whole conformational space, so that we can determine how side-chain excluded volume affects protein stability. No other type of model is yet available that can tell us about such conformational spaces in an exact way.

Packing density

We define packing density in 2 different ways for 2 different purposes. For the purpose of comparing the configurational freedom of the SCM and the LCM, we define the density in terms of the number of *main-chain* monomers per unit volume. Because we can compare SCM and LCM configurations with the same main-chain configuration, main-chain packing density provides a parameter for studying the configurational freedom and excluded volume effects of appended side chains. But for studying the distribution of contacts within the SCM, we define density in terms of the *total* number of monomers (main chain plus side chain) per unit volume, because this definition gives more interpretable limiting values when we consider the SCM alone.

We use 2 dimensionless parameters for the packing density, the number of monomers per unit volume for a particular configuration divided by the number of monomers per unit volume of a maximally compact configuration. They are: (1) ρ_s , a function of R_s , the radius of gyration of all monomers, main chain plus side chain; and (2) ρ_l , a function of R_l , the radius of gyration of the main-chain monomers alone (see Methods). The difference between ρ_s and ρ_l is illustrated in Figure 2, which shows identical main-chain configurations with side chains (SCM) and without them (LCM). $\rho_l = 0.33$ for both cases. For the LCM configuration, $\rho_l = \rho_s$ because all the monomers are main-chain monomers. For the SCM configuration, $\rho_s = 0.57 > \rho_l$ because the side chains contribute to R_s and ρ_s .

Richards (1974) has defined packing density in proteins as "the ratio of the volume enclosed by the van der Waals envelope of a given molecule to the actual volume of space that it occupies." Recently, to include lattice and other model studies, Richards and Lim (1994) have revised this definition to be "the ratio of the actual volume in 3D, or area in 2D, of the object to the volume or area that it occupies in the assembly under consideration." For crystals of simple organic compounds and for the interiors of proteins, the packing density by this definition is about 0.74 (Richards, 1974).

Lattice models cannot address variations of packing density due to voids of atomic volume, the subject of Richards's seminal studies (Richards, 1974, 1977). However, our focus is on conformational changes and protein folding. In protein/water systems at 1 atmosphere pressure, conformational changes are not accompanied by significant changes in free volume (Richards, 1974, 1977). Therefore our lattice representation of density describes the relevant changes in model proteins as they expand their radii of gyration in solution and as protein lattice sites become interspersed with solvent lattice sites. To complete the correspondence with the first definition of packing density due to Richards, we could simply define each monomer as occupying 0.74 of its corresponding lattice site (see Fig. 12). However, this factor drops out in our calculation of packing density (see Equation 10 below), hence it has no bearing on our present study.

Results

Contributions of side chains to the entropy of polymer collapse

How does the introduction of side chains affect the distribution of configurations and the entropy of folding? To address this

question, we compare the distributions of the SCM and LCM configurations for chains of a given chain length, n = the number of main-chain monomers, over ρ_l . The fraction of SCM configurations that are found in the i th 10% interval of ρ_l , where $0 < \rho_l < 1$, $f_S(\rho_l)$, is given by

$$f_S(\rho_l) = \frac{\Omega_S(\rho_l(i))}{\sum_{i=1}^{10} \Omega_S(\rho_l(i))}, \quad (1)$$

where $\Omega_S(\rho_l(i))$ is the number of SCM configurations found in the i th interval of ρ_l .

The fraction of LCM configurations that are found in the i th 10% interval of ρ_l is given by

$$f_L(\rho_l) = \frac{\Omega_L(\rho_l(i))}{\sum_{i=1}^{10} \Omega_L(\rho_l(i))}, \quad (2)$$

where $\Omega_L(\rho_l(i))$ is the number of LCM configurations found in the i th interval of ρ_l .

Figure 3 shows the distributions of f_L and f_S over ρ_l . As predicted by scaling arguments (deGennes, 1979; Freed, 1987), the distributions tend to peak at lower density as chain length and dimensionality increase. In both 2D and 3D, LCM and SCM chains have low average density when the only energetic interactions are due to excluded volume (as in a good solvent).

We define the excess entropy due to side chains, $\Delta S/nk$, using the ratio of $f_S(\rho_l)/f_L(\rho_l)$ for all conformations of density ρ_l and chain length n main-chain monomers:

$$\frac{\Delta S(\rho_l)}{nk} = \ln \left[\frac{f_S(\rho_l)}{f_L(\rho_l)} \right], \quad (3)$$

where k is Boltzmann's constant. $\Delta S(\rho_l)/nk$ versus ρ_l is shown in Figure 4. Side chains increase the number of degrees of freedom and the conformational entropy in chains with open main-chain conformations: $\Delta S(\rho_l)/nk > 0$. (See Figure 5A, which shows all the possible orientations of side chains on a 2D open

Fig. 3. Distributions of configurations $f(\rho_l)$ versus main-chain density ρ_l for (A) 2D and (B) 3D for the SCM (---□---) and LCM (—■—) determined by exhaustive simulations for $n = 16$ in 2D and $n = 10$ in 3D and by Rosenbluth weighted MC sampling for $n = 50$. Error bars in this figure (which are smaller than the points) and in Figures 4, 6, and 7 indicate 1 standard error of the mean values from 5 MC runs of 1.7×10^7 configurations.

Fig. 4. Excess entropies due to appending side chains, $\Delta S(\rho_l)/nk$, versus ρ_l in (A) 2D and (B) 3D determined by exhaustive (—) simulations for $n = 16$ in 2D and $n = 10$ in 3D and by Rosenbluth weighted MC (---) sampling for $n = 50$.

main-chain configuration. For this open configuration, only a few combinations of side-chain orientations are forbidden by steric conflicts.) But for chains that are more compact, depending on chain length and dimensionality, we find that $\Delta S(\rho_l)/nk < 0$, indicating that the steric constraints due to the pendant side-chain volumes reduce the conformational freedom. (See Figure 5B, which shows a compact main-chain configuration forbidden in the SCM due to the lack of an available site for the addition of one of the side chains.) In both 2D and 3D, more conformational freedom is lost upon compaction due to side-chain excluded volume as the chain length increases.

Figure 4 shows that for longer chains the side-chain entropy is a steep function of packing density near the maximally com-

Fig. 5. A: Side chains add degrees of freedom to open main-chain configurations. In 2D, each side chain may occupy 1 of 2 possible sites (indicated by arrows) adjacent to the main-chain monomer (■) to which it is attached, if the possible site is not occupied by a main-chain or another side-chain monomer. Certain combinations of side-chain placements are forbidden by steric conflicts, indicated here by the overlapping arrows. B: Side chains add steric hindrance that decreases the number of possible compact main-chain configurations. Here the main chain is compact, and all the sites adjacent to the main-chain monomer marked X are occupied by other main-chain monomers. It is impossible to add a side-chain to main-chain monomer X so this main-chain configuration is forbidden in the SCM.

compact state. This result suggests that “side-chain freezing” (Dill, 1990) may occur in protein folding; that is, much side-chain freedom may be lost, or frozen out, in the last stages of folding. Conversely, only a small volume expansion from the maximally compact states leads to a large increase in side-chain freedom. Hence this model shows no critical disjuncture point.

The side chains in the SCM reduce the fraction of compact states relative to the LCM, but they cannot discriminate one maximally compact state from another. That is, because side chains are identical in the SCM, this model has no steric complementarity that could be construed to encode the distinction between 2 different native conformations. Nevertheless, the side chains in the SCM significantly restrict the ways the entire chain can pack. We expect the effects of different side-chain sizes and shapes on the packing of side chains to be perturbations to this simple model.

The distributions of contacts shown in “How do the frequencies of contact types change with packing density?” below are given as a function of ρ_s , because we wish to understand how the patterns of contacts change as the SCM becomes maximally compact. In the LCM, $\rho_s = \rho_l$, and $f_L(\rho_s) = f_L(\rho_l)$. On the other hand, in the SCM, $f_S(\rho_s)$ is given by

$$f_S(\rho_s) = \frac{\Omega_S(\rho_s(i))}{\sum_{i=1}^{10} \Omega_S(\rho_s(i))}, \quad (4)$$

where $\Omega_S(\rho_s(i))$ is the number of SCM configurations found in the i th interval of ρ_s . In Figure 6 we show the distribution of LCM and SCM configurations versus ρ_s . Considering the full ensemble, adding side chains leads to higher average density. Nonetheless, due to the steric exclusion of the side chains, the SCM has a smaller fraction of configurations at maximum ρ_s than the LCM.

Patterns of intrachain contacts

When a polymer is modeled as a linear string of beads, it can have only 1 type of intrachain contact: a (main-chain) monomer may contact another (main-chain) monomer. But the SCM has 2 types of monomer, main chain and side chain, so there are

Fig. 6. Distributions of configurations $f(\rho_s)$ versus total density ρ_s in (A) 2D and (B) 3D for the SCM (---□---) and LCM (—■—) determined as described in Figure 3.

3 categories of topological contacts (t): main chain to main chain (mm), side chain to side chain (ss), and main chain to side chain (ms). Contacts in each category can be local or non-local with respect to sequence position, as defined in Figure 2. In this section, we explore the distributions of contacts over the different types, and over the range of packing densities.

In order to compare the numbers of intrachain contacts in configurations of lattice models with different topologies and dimensionalities, we use the normalizing parameter $t_{\max}$, the maximum possible number of intrachain contacts for a chain of a given number of monomers. (Chan and Dill [1989b, 1990b] defined $t_{\max}$ for the 2D and 3D LCM.¹) For any category of contact i = (local or nonlocal t , mm , ss , ms), the fraction Φ_i of the $t_{\max}$ contacts belonging to type i is given by

$$\Phi_i = \frac{n_i}{t_{\max}}, \quad (5)$$

where n_i is the number of type i contacts in a given conformation. Ensemble averages of Φ_i over all configurations are denoted $\langle \Phi_i \rangle$. Averages over 10% intervals of density are denoted $\langle \phi_i \rangle$.

Averages of contact types

Figure 7 shows that, in both 2D and 3D, (1) most contacts are local and (2) ms and ss contacts are much more prevalent than mm contacts. However, the 2D and 3D results differ in the following respects: (1) Contacts, especially those that are local, are more prevalent in 2D than in 3D. (2) Whereas ss contacts are slightly more common than ms contacts in 2D, the relationship is reversed in 3D. Here we explore the factors that determine the frequencies of the contact categories and lead to the rank ordering $ss > ms \gg mm$ in 2D and $ms > ss \gg mm$ in 3D.

The trends in contact frequencies are given qualitatively, but not quantitatively, by the following simple argument. In 2D, 1/4 of the square main-chain monomer sides and 3/4 of the square side-chain monomer sides are not involved in covalent links and are therefore available to form topological contacts. Thus, neglecting the effects of chain interference, the a priori probabilities of contacts are predicted to be $p(mm) = (1/4)(1/4)$, $p(ss) = (3/4)(3/4)$, and $p(ms) = 2(1/4)(3/4)$ in 2D. The factor of 2 in $p(ms)$ accounts for the 2 different arrangements of ms contacts: an ms contact between monomers i and j ($i < j$) can occur between main-chain monomer i and side-chain j or side-chain i and main-chain monomer j (see Fig. 8). This simple argument predicts the ratios $p(mm):p(ss):p(ms)$ to be 1:9:6 in 2D. By the same logic, the ratio 1:2.8:3.3 is predicted on the 3D simple cubic lattice.

The simulations show that, when chain connectivity is taken into account, the values of the contact probability ratios for long chains in the ensemble average are

$$\text{2D: } \langle \Phi_{mm} \rangle : \langle \Phi_{ss} \rangle : \langle \Phi_{ms} \rangle = 1:22:15, \quad \text{and}$$

$$\text{3D: } \langle \Phi_{mm} \rangle : \langle \Phi_{ss} \rangle : \langle \Phi_{ms} \rangle = 1:4.25:5. \quad (6)$$

¹ These definitions are readily extended to estimate $t_{\max}$ in the SCM by substituting $\nu = 2n - 2$, the number of monomers in an SCM chain of $N + 1$ main-chain monomers, for $N + 1$, the number of monomers in an LCM chain of N bonds.

Fig. 7. Averages of contact types. Local and nonlocal $\langle \Phi_i \rangle$ are determined by Rosenbluth weighted MC sampling for chains with $n = 50$ main-chain monomers in 2D and 3D. The shaded regions indicate local contacts and the unshaded regions indicate nonlocal contacts. Standard errors of the means, determined as described in the caption to Figure 3, are all less than $(1.2 \times 10^{-3})t_{\max}$.

Hence, interference by other parts of the chain affects the contact probabilities.

Figures 8 and 9 illustrate how the rest of the chain interferes with the formation of the different types of local contacts. A local contact restricts the conformations of the 3 bonds in the

Fig. 8. Definitions of specified contacts for studying chain interference on local *ss*, *ms*, and *mm* contacts with equal length tails. ■ indicates a main-chain monomer in a contact loop, and □ indicates a side-chain monomer in or attached to a local contact loop in the 2D SCM. The number of arrows leading to or from a contact loop indicates the number of possible steps for main-chain monomers $i-1$ or $j+1$. **A:** A local *ss* contact loop is comprised of 2 side-chain steps and 1 main-chain step; n is even. **B:** A local *mm* contact loop is comprised of 3 main-chain steps; n is even. **C:** A local *ms* contact loop is comprised of 1 side-chain step and 2 main-chain steps; n is odd.

loop connecting the main-chain or side-chain monomers i and j participating in the contact. We show how the rest of the chain can interfere with a contact by computing how the probability of a specified contact depends on the equal lengths l of 2 “tail” chains, attached to main-chain monomers i and j . The number of main-chain monomers in the chain is $n = 2l + c + 1$, where c is the number of main-chain bonds in the loop (see Fig. 8). We compare the reduction factors (Chan & Dill, 1989a, and references therein) $R_S(n, i)$ for $i = (\text{local}, mm, ss, \text{or } ms)$ contacts for the SCM, and $R_L(n, mm)$ for the LCM, given by

$$R_S(n, i) = \frac{\Omega_S(n, i)}{\Omega_S(n)}, \quad \text{and}$$

$$R_L(n, mm) = \frac{\Omega_L(n, mm)}{\Omega_L(n)}, \quad (7)$$

where $\Omega_S(n, i)$ is the number of SCM configurations of chain length n that contain the specified local contact, $\Omega_L(n, mm)$ is the number of LCM configurations of chain length n that contain the specified local contact, $\Omega_S(n)$ is the total number of SCM configurations, and $\Omega_L(n)$ is the total number of LCM configurations of a given chain length n . Smaller values of R indicate that a smaller fraction of all configurations contain the specified contact.

Figure 9 shows how the contact reduction factors depend on the chain length. (Because there are twice as many ways for a given pair of model residues to form an *ms* contact as *mm* or *ss* contacts, we show $R_S(n, ms)/2$ in Figure 9.) Initial increases in the length of the 2 equal tails generally increase the restrictions on forming local contacts (exceptions are discussed below). After adding a few tail monomers, however, there is little additional interference.

The different types of local contacts impose different degrees of conformational restriction. Local contacts between side-chain

Fig. 9. Reduction factors, $R_S(n, i)$ and $R_L(n, mm)$, indicating how chain tails interfere with the formation of local *ss*, *ms*, and *mm* contacts with equal length tails, were determined by exhaustive simulation for chain lengths $4 \leq n \leq 16$ in 2D and chain lengths $4 \leq n \leq 10$ in 3D and by Rosenbluth weighted MC sampling for $20 \leq n \leq 55$ in 2D and 3D. Error bars indicate the standard error of the mean determined from 3 MC runs of 12×10^5 configurations.

monomers (*ss*) are less costly in chain entropy than local contacts between main-chain *i* and side-chain *j* monomers (*ms*), which are less costly than local contacts between 2 main-chain monomers (*mm*). In fact, side-chain monomer *i* in a local contact can encounter no steric conflict with monomers 2 bonds away from the contacting monomer *j*, whereas side-chain monomers that are not in a local contact can encounter such steric conflict. Thus, the number of configurations containing a specified local *ss* contact is a larger proportion of the total number of configurations for chain length $n = 6$ than for chain length $n = 4$, and $R_S(6, ss) > R_S(4, ss)$. For the same reasons $R_S(7, ms) > R_S(5, ms)$. These details are dependent on the bond angles defined by the lattice. However, for both local and nonlocal contacts, the main differences between the main-chain and side-chain contacts are independent of bond angles: (1) more of the side-chain surface than the main-chain surface is available for making contacts; (2) main-chain contacts directly restrict the path of the main chain, as shown in Figure 8; and (3) the restriction due to main-chain contacts is more severe when side chains are appended.

For local contacts in the middle of a polypeptide chain, these *R* factors suggest that *ss* contacts among side chains may be less costly in chain entropy than helix capping main-chain-side-chain *ms* hydrogen bonds, which may be less costly than main-chain *mm* hydrogen bonds. The Φ_i given Figure 7, which are averages over all loop lengths and sequence positions, suggest that the differences among the contact types may hold generally: that the configurational entropy costs of contacts with a given loop length and sequence position may be greater in general for main-chain contacts than for side-chain contacts. But because the study of *R* factors shows that local lattice details contribute to the differences among contact categories, and because we have investigated only the *R* factors for local contacts in the middle of the chain sequence, further studies are needed to determine the generality of the difference among the contact types with respect to the model, loop length, and sequence position.

How do the frequencies of contact types change with packing density?

Figure 10 shows that (1) in both 2D and 3D, the frequencies of *ss*, *ms*, and *mm* contacts increase as density increases, but at different rates; (2) in 3D the increase in nonlocal $\langle \phi_{ms} \rangle$ leads the increase in contacts over all densities; (3) the increase in $\langle \phi_{mm} \rangle$ is a large factor multiplied by a small number, and *mm* contacts do not contribute more than 10% of $t_{\max}$ until the chains are quite compact; (4) local $\langle \phi_{ms} \rangle$ and $\langle \phi_{ss} \rangle$ are generally flat over density, but local *mm* contacts depend on density. Because *mm* contacts correspond to protein main-chain hydrogen bonds, the implication is that protein main-chain hydrogen bonding becomes more favored as the protein becomes compact. This result is not inconsistent with reports of main-chain amide exchange protection detected by NMR early in the folding pathway. Because compactness is achieved early in folding pathways, during the dead time of experiments (Roder et al., 1988; Miranker et al., 1991; Radford et al., 1992), the enhanced hydrogen bonding due to compactness should also occur early. Moreover, as Kim et al. (1993) point out, amide hydrogen exchange protection in denatured states does not necessarily imply the presence of hydrogen bonds but indicates only that the amide proton is less accessible for exchange. Rapid collapse may cause steric hindrance to main-chain amide hydrogen exchange.

Fig. 10. Distributions of contact types versus chain packing density. **Upper panel:** Bars show SCM, total of all contact types; line (—) shows LCM all contacts = $\langle \phi_{mm} \rangle$ because there are only *mm* contacts in the LCM; dotted line (.....) shows only local $\langle \phi_{mm} \rangle$ in the LCM. **Lower panels:** $\langle \phi_{ms} \rangle$, $\langle \phi_{ss} \rangle$, and $\langle \phi_{mm} \rangle$ in the SCM. Shaded regions indicate local contacts; unshaded regions indicate nonlocal contacts. Results are determined by Rosenbluth weighted MC simulations for chain length $n = 50$ in (A) 2D and (B) 3D in 5 MC runs of 1.7×10^7 configurations. The absolute standard errors for all quantities are less than $0.02t_{\max}$.

In contrast, local *ms* contacts, corresponding to protein main-chain to side-chain contacts including helix capping hydrogen bonds (Stickle et al., 1992), are common even at low density in our simulations. These results are consistent with the suggestions that hydrogen bonds between main chain and side chains may occur as open chains begin to fold in protein α -helices (Presta & Rose, 1988; Richardson & Richardson, 1988; Bruch et al., 1991; Lyu et al., 1993; Rose & Wolfenden, 1993).

Comparisons with the PDB

We find that the distributions of *mm*, *ms*, and *ss* contacts in compact SCM configurations resemble those of proteins in the PDB (Bernstein et al., 1977; Abola et al., 1987). Figure 11 compares weighted distance maps (Yee & Dill, 1993) for 50 different randomly generated 3D SCM configurations of length $n = 40$ and densities $0.9 \leq \rho_s \leq 1.0$ with the maps for 20 crystallographically determined protein configurations from the PDB. The data set of PDB proteins, given in Table 1, is the subset of proteins less than 100 amino acids long, from the set of proteins that were chosen as representative of the PDB by Lesk (1991).

The weighted distance map is determined as follows. For each lattice or PDB configuration, contact weights are computed for each pair of amino acids or lattice monomers *i* and *j* that are not adjacent in the sequence of the chain. The contact weight w_{ij} is defined by: $w_{ij} = d_{ij}^{-6}$, where d_{ij} is the distance between units *i* and *j*. Thus, the pairs that are close in space are assigned

Fig. 11. The frequencies of contacts, f_{oi} , as a function of contact order are computed for 20 protein configurations from the Brookhaven PDB (top row) and 50 different randomly generated Rosenbluth weighted 3D SCM lattice configurations of chain length $n = 40$ with densities $0.9 \leq \rho \leq 1.0$ (bottom row).

large contact weights, whereas the pairs that are distant in space are assigned small contact weights. The use of contact weights eliminates the need for an arbitrary cutoff distance to define a contact so that contact frequencies for lattice model and real protein configurations can be compared.

We compare distributions of contact types in the SCM and PDB structures as follows. The contact order is defined (1) for the lattice model as the number of main-chain bonds separating the contacting monomers and (2) for proteins as the number of virtual C_α to C_α bonds separating the contacting amino acids. The contact frequency of order o and type i , f_{oi} , is defined as the sum of all type $i = (mm, ms, \text{ or } ss)$ contact weights of order o divided by the total type i contact weights of the configurations in each data set (see Methods). By defining contact frequency in this way, we compare only the relative frequencies of the different categories of contacts in the PDB and in the SCM. The factor of $t_{\max}$ that was used for comparisons among lattice models does not enter this calculation. The ratios of the total type i contact weights of the configurations in the SCM data set, $T_i(\text{SCM})$, are similar to those of the PDB data set, $T_i(\text{PDB})$:

$$T_{mm}(\text{SCM}) : T_{ms}(\text{SCM}) : T_{ss}(\text{SCM}) = 1 : 1.7 : 1.1, \quad \text{and}$$

$$T_{mm}(\text{PDB}) : T_{ms}(\text{PDB}) : T_{ss}(\text{PDB}) = 1 : 1.9 : 1.4.$$

For the total PDB data set compiled by Lesk (1991), including longer proteins, the corresponding ratio is 1:1.9:1. The large contribution from ms contacts may be significant in protein stability. Previous studies have reported on the ms contacts in proteins that are hydrogen bonds. Stickle et al. (1992) observed in a survey of 45 proteins from the PDB that 21% of all hydrogen bonds occur between main chain and side chains. Earlier, in a data set of 15 proteins, Baker and Hubbard (1984) observed that 11% of all main-chain carbonyl oxygens and amide nitrogens hydrogen bond to side-chain atoms.

Figure 11 shows that the 3D SCM captures several important features of the distributions of contacts over contact order in the PDB. (1) Most contacts have low contact order, i.e., they involve monomers i and j close together in the sequence. (2) The distributions of mm contacts are the most strongly skewed to

Table 1. References for the PDB protein data set (Fig. 11)

451c	Matsuura Y, Takano T, Dickerson RE. 1982. Structure of cytochrome c_{551} from <i>P. aeruginosa</i> refined at 1.6 Ångströms resolution and comparison of the two redox forms. <i>J Mol Biol</i> 156:389–409.
2abx	Love RA, Stroud RM. 1986. The crystal structure of alpha-bungarotoxin at 2.5 Ångströms resolution. Relation to solution structure and binding to acetylcholine receptor. <i>Protein Eng</i> 1:37–46.
1cc5	Carter DC, Melis KA, O'Donnell SE, Burgess BK, Furey WF Jr, Wang BC, Stout CD. 1985. Crystal structure of <i>Azotobacter</i> cytochrome c_5 at 2.5 Ångströms resolution. <i>J Mol Biol</i> 184:279–295.
1crn	Teeter MM. 1984. Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. <i>Proc Natl Acad Sci USA</i> 81:6014–6018.
1ctf	Leijonmarck M, Liljas A. 1987. Structure of the C-terminal domain of the ribosomal protein L7/L12 from <i>Escherichia coli</i> at 1.7 Ångströms. <i>J Mol Biol</i> 195:555–579.
1ctx	Walkinshaw MD, Saenger W, Maelicke A. 1980. Three-dimensional structure of the long neurotoxin from cobra venom. <i>Proc Natl Acad Sci USA</i> 77:2400–2404.
5ebx	Corfield PWR, Lee TJ, Low BW. 1989. The crystal structure of erabutoxin a at 2.0 Ångströms resolution. <i>J Biol Chem</i> 264:9239–9342.
1fdx	Adman ET, Sieker LC, Jensen LH. 1976. Structure of <i>Pep-tococcus aerogenes</i> ferredoxin, refinement at 2 Ångströms resolution. <i>J Biol Chem</i> 251:3801–3806.
1fxc	Tsukihara T, Fukuyama K, Nakamura M, Katsube Y, Tanaka N, Kakudo M, Wada K, Hase T, Matsubara H. 1981. X-ray analysis of a (2^*Fe-2^*S) ferredoxin from <i>Spirulina platen-sis</i> . Main chain fold and location of side chains at 2.5 Ångströms resolution. <i>J Biochem (Tokyo)</i> 90:1763–1773.
2gn5	Brayer GD, McPherson A. 1983. Refined structure of the gene 5 DNA binding protein from bacteriophage FD. <i>J Mol Biol</i> 169:565–596.
1hip	Carter CW Jr, Kraut J, Freer ST, Xuong NH, Alden RA, Bartsch RG. 1974. Two Ångstrom crystal structure of oxidized chromatin high potential iron protein. <i>J Biol Chem</i> 249:4212–4225.
3icb	Szebenyi DMW, Moffat K. 1986. The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. <i>J Biol Chem</i> 261:8761–8777.
1nxb	Tsernoglou D, Petsko GA, Hudson RA. 1978. Structure and function of snake venom curarimimetic neurotoxins. <i>Mol Pharmacol</i> 14:710–716.
1ovo	Papamokos E, Weber E, Bode W, Huber R, Empie MW, Kato I, Laskowski M Jr. 1982. Crystallographic refinement of Japanese quail ovomucoid, a Kazal-type inhibitor, and model building studies of complexes with serine proteases. <i>J Mol Biol</i> 158:515–537.
2ovo	Bode W, Epp O, Huber R, Laskowski M Jr, Ardelt W. 1985. The crystal and molecular structure of the third domain of silver pheasant ovomucoid (OMSVP3). <i>Eur J Biochem</i> 147:387–395.
1pcy	Guss JM, Freeman HC. 1983. Structure of oxidized poplar plastocyanin at 1.6 Ångströms resolution. <i>J Mol Biol</i> 169:521–563.
5pti	Wlodawer A, Walter J, Huber R, Sjolín L. 1984. Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. <i>J Mol Biol</i> 180:301–329.
5rxn	Watenpaugh KD, Sieker LC, Jensen LH. 1980. Crystallographic refinement of rubredoxin at 1.2 Ångströms resolution. <i>J Mol Biol</i> 138:615–633.
1sn3	Almasy RJ, Fontecilla-Camps JC, Suddath FL, Bugg CE. 1983. Structure of variant-3 scorpion neurotoxin from <i>Centruroides sculpturatus</i> Ewing, refined at 1.8 Ångströms resolution. <i>J Mol Biol</i> 170:497–527.
1ubq	Vijay-Kumar S, Bugg CE, Cook WJ. 1987. Structure of ubiquitin refined at 1.8 Ångströms resolution. <i>J Mol Biol</i> 194:531–544.

low contact order, followed closely by *ms*, and then trailed by *ss* contacts. These 2 properties hold for longer chains in the PDB as well. (3) The frequency of the lowest order *mm*, *ss*, and *ms* contacts in the 3D SCM are within 2% of the corresponding frequencies of lowest order contacts in the short-chain PDB data set. However, the randomly generated $n = 40$ compact lattice-chain configurations show an excess of middle-order contacts. This is partly due to the greater chain length of the short-chain PDB data set. The mean chain length of the short-chain PDB data set is 71 amino acids and contacts of order greater than 40 account for over 13% of its total contact weight. As expected, the contact frequencies of long 2D SCM chains are more skewed to the local side than those of the 3D SCM or the PDB.

Views of packing in proteins

The SCM goes 1 step beyond simple polymer linear bead models in realism and predicts a large side-chain excluded volume entropy that opposes protein folding. But is this side-chain model still too simple to capture the essential side-chain packing or thermodynamic properties of proteins? We believe that the present model gives a reasonable first approximation for the role played by side chains. In this section, we consider a range of experiments that bear on side-chain effects. For an excellent comprehensive review of side-chain packing in proteins see Richards and Lim (1994).

On the one hand, the ability to achieve the tight packing characteristic of native proteins can arise in either jigsaw puzzle or nuts and bolts models such as the SCM (see Fig. 12). On the other hand, there is also evidence for considerable "flexibility" in protein structures that appears to be more consistent with nuts and bolts than jigsaw puzzle models. For example, protein families such as the globins can achieve the same fold with little sequence similarity (Bowie et al., 1991; Richards & Lim, 1994, and references cited therein). Moreover, proteins are tolerant of change and retain their overall native fold even when mutations lead to considerable changes in core side-chain volumes (Matthews, 1987; Lim & Sauer, 1989; Bowie et al., 1990; Keefe et al., 1993; Richards & Lim, 1994). McRee et al. (1990), Eriksson et al. (1992), and Varadarajan and Richards (1992) have found

Fig. 12. The SCM can collapse to high density and fill space without voids. Alternatively, maximum density may be regarded as 0.74 as in proteins because, as shown here, monomers need not be square in 2D or cubic in 3D.

that many cavities in protein cores are not rigid like jigsaw puzzles that lack a piece. Rather, the protein structure may "relax" to accommodate mutations. Baldwin et al. (1993) have also recently shown that main-chain readjustments accompany side-chain volume changing core mutations. Computer simulations by Novotny et al. (1984) suggest that native compactness in protein structures is readily achieved even by misfolded sequences. There is further evidence in protein dynamics studies for the flexibility of protein interiors: small molecules rapidly diffuse to buried residues (Gurd & Rothgeb, 1979; Karplus & McCammon, 1981), and aromatic rings rotate with low activation energy barriers (Wüthrich & Wagner, 1978). Finally, studies of protein native structures show that neighboring core side chains have little or no tendency toward specific preferred mutually buried areas (Behe et al., 1991), and little preference for particular relative orientations of spatially neighboring hydrophobic side chains (Singh & Thornton, 1990, 1992). In these respects side-chain organization does not resemble the packing of a jigsaw puzzle.

To what extent does side-chain packing contribute to specifying the native fold of a protein? Even if all side chains have uniform size and shape, as in the SCM, they contribute to specifying a native fold by reducing the number of viable compact conformations. The drive to compact states reduces the number of conformations by many orders of magnitude and in this respect contributes to selecting the native fold from among all possible conformations (Dill, 1985; Chan & Dill, 1990a). Figures 3, 4, 5, and 6 show that the addition of uniform side chains to a linear chain model further selects from among the possible compact states those configurations that can also accommodate the side chains. Excluded volume is a factor in native state selection, and side-chain packing contributes to it.

But the steric hindrance of main chain and side chains does not further discriminate among the many viable compact states. Thus, the steric forces alone that are encoded in the excluded volume of the present model cannot discriminate the lysozyme fold from the ribonuclease fold because every side chain is identical to every other and there are no differences among sequences. On the other hand, it has been shown that many sequences of nonpolar and polar monomers can fold with a high degree of discrimination among compact states, even in a linear chain model. When nonpolar monomers are interspersed with polar monomers in specific heteropolymeric sequences, linear heteropolymeric chains have very few conformations of lowest energy, sometimes just one (Dill, 1985; Lau & Dill, 1989; Chan & Dill, 1991; Camacho & Thirumalai, 1993). Recent experiments with model proteins support this view. Kamtekar et al. (1993) have designed only hydrophobic and polar aspects into amino acid sequences and appear to have constructed 4-helix bundles, without designing the side-chain packing, although it is not yet clear whether these structures are unique like native proteins. Hence, at the level of these models, the sequence of hydrophobic and polar monomers contributes more to dictating the chain fold than side-chain sizes or shapes.

Is side-chain packing the basis for the cooperativity in protein folding? Shakhnovich and Finkelstein (1989) have proposed that it is. According to their model (the SF model), a sudden freezing of side-chain rotations is the origin of the cooperativity of protein folding. The SF model requires a 25% expansion in volume relative to the native core volume before a sudden unfreezing of core side-chain rotameric degrees of freedom incurs a sharp increase in entropy. The main chain is maintained in a rigid native-like conformation, except at hinge points that al-

low expansion. A 25% expansion represents an added volume the equivalent of at least a dozen side chains in the core, adding about 40 side-chain volumes throughout a whole protein of 100 amino acids. The cooperativity in the SF model requires a critical disjuncture point, characteristic of jigsaw puzzle packing, to give rise to a plateau and consequent sharp increase in the entropy as density decreases, and to a barrier in the free energy versus density. The SF model is the basis for the definition of the "molten globule" (Shakhnovich & Finkelstein, 1989; Karplus & Shakhnovich, 1992; Ptitsyn, 1992) as the state in which a protein has fixed native-like main chain and freely rotating side chains.

The SCM leads to a rather different picture of entropic contributions to cooperativity. Figure 13 compares the side-chain entropies in jigsaw puzzle models, such as the SF model, and nuts and bolts models, such as the SCM. The 2 models predict that side-chain rotational entropy is lost at very different densities. In the SCM, a small expansion in volume from the compact native state produces a *steep increase* in side-chain rotational entropy. The increase in side-chain degrees of freedom is linked to the increase in main-chain degrees of freedom. In the SF model, the same expansion leads to *no change*, i.e., a plateau in side-chain rotational entropy.

Neither the SF model nor the SCM draws definitive conclusions about the origins of cooperativity in protein folding. On the one hand, the SF model has many approximations and assumptions. On the other hand, the SCM is a low resolution model. It neglects internal side-chain degrees of freedom, the different sizes and shapes of the amino acids, and all energetic interactions except excluded volume. It is not yet known whether side-chain degrees of freedom are an important component of protein folding cooperativity. The present work simply suggests that side chains may contribute considerable entropy opposing protein stability, without being the principal basis for the cooperativity of folding. Another significant distinction is that the SF model predicts that folding cooperativity results even while the main chain remains in a fixed native-like configuration. In contrast, the side-chain freedom is inextricably correlated with the freedom of the main chain in the SCM.

Fig. 13. A schematic comparison shows the qualitative difference in the dependencies of side-chain entropy on chain density in: (A) a jigsaw puzzle model for side-chain packing, based on the SF model for the torsional motion of aliphatic side chains in which main-chain degrees of freedom do not contribute and the increase in entropy at high density is due to vibrations and not to side-chain rotation; and (B) a connected chain nuts and bolts model such as the SCM in which main-chain and side-chain degrees of freedom are linked.

Conclusions

We have explored the degrees of freedom and the patterns of intrachain contacts in a simple polymer lattice model and we have found that side-chain entropies are large and should be taken into account in modeling protein stability. Side chains in this simple model of nuts and bolts packing (1) strongly oppose folding, compared to the linear chain; (2) only freeze completely into place and lose much entropy in the final approach to maximum compactness; (3) give rise to different contact frequencies among and between side chains and main chain in open states; (4) give rise to distributions of the different contact types in compact states similar to those observed in the PDB; (5) have no steric complementarities and therefore predict tolerance to mutations and relatively random distributions of mutually buried areas as observed by Behe et al. (1991); and (6) do not by themselves lead to an entropy barrier that could account for the 2-state thermodynamics of protein folding. The SCM is simple enough to permit explorations of open conformations without approximation, including main-chain degrees of freedom, and to predict how configurational freedom and patterns of intrachain contacts depend on chain density.

This simple nuts and bolts model of uniform side chains shows how side chains can contribute a large entropy opposing native stability, even in the absence of locking interactions. With this information in hand, it would be fruitful to test more detailed models, including off-lattice models, to account for the additional properties that may arise from the various sizes and shapes of amino acid side chains.

Methods

We exhaustively enumerated LCM and SCM self-avoiding configurations up to $n = 16$ main-chain monomers long in 2D, and up to $n = 10$ main-chain monomers long in 3D. LCM and SCM configurations of chains up to $n = 50$ or 55 main-chain monomers long in both 2D and 3D were generated by MC sampling using the configuration weighting method of Rosenbluth and Rosenbluth (1955) to estimate conformation probabilities and distribution entropies (Meirovitch, 1982). Simulations were run on a Sun SPARCstation I workstation and HP-730 computer. The representation of side-chain degrees of freedom was given by unoccupied sites adjacent to the main-chain monomers: a maximum of 2 side-chain degrees of freedom in 2D and a maximum of 4 side-chain degrees of freedom in 3D.

For each configuration, R_s , the radius of gyration of all the monomers, and R_l , the radius of gyration of all main-chain monomers alone, was computed. R_s^2 is given by

$$R_s^2 = \frac{1}{\nu} \sum_{i=1}^{\nu} (\mathbf{r}_i - \langle \mathbf{r} \rangle)^2, \quad \langle \mathbf{r} \rangle = \frac{1}{\nu} \sum_{i=1}^{\nu} \mathbf{r}_i, \quad (8)$$

where ν is the total number of monomers in the chain. In the LCM, $\nu = n$, the number of main-chain monomers. In the SCM, $\nu = 2n - 2$. (Defining the chain ends as unbranched is one way to avoid ambiguity in defining the chain length, n . Other alternatives are possible, but the importance of this choice diminishes as n increases.) $\mathbf{r}_i$ is the position of the i th monomer and $\langle \mathbf{r} \rangle$ is the centroid of the configuration.

R_I^2 is given by

$$R_I^2 = \frac{1}{n} \sum_{i=1}^n (\mathbf{r}_i - \langle \mathbf{r}_I \rangle)^2, \quad \langle \mathbf{r}_I \rangle = \frac{1}{n} \sum_{i=1}^n \mathbf{r}_i, \quad (9)$$

where the index i runs over all main-chain monomers, which contribute to R_I in both the SCM and LCM ($R_S = R_I$ in the LCM). The dependence of the mean square radii of gyration $\langle R_S^2 \rangle$ (the brackets denote the ensemble average), Ω_L , and Ω_S on chain length is shown in Table 2.

For configuration j , $\rho_s(j)$ and $\rho_l(j)$ are given by:

$$\rho_s(j) = \left[\frac{R_{s0}}{R_s(j)} \right]^d, \quad \rho_l(j) = \left[\frac{R_{l0}}{R_l(j)} \right]^d, \quad (10)$$

where d is the number of dimensions (2 or 3 in the present work), R_{s0} and R_{l0} are the minimal radii of gyration, corresponding to the maximally compact conformations, and $R_s(j)$ and $R_l(j)$ are the radii of gyration of configuration j . Thus, ρ_s is the packing density of the entire chain including side chains and ρ_l is the packing density of the main chain alone.

R_{l0} and R_{s0} for short chains were determined by exhaustive simulation. For chains too long for exhaustive enumeration R_{s0}

is determined by the following approximations. R_{s0} is approximated when ν is a perfect square in 2D or a perfect cube in 3D. For much longer lattice chains, minimum radii of gyration are found in configurations that are approximately circular or spherical. But for the chain lengths explored here, the rectilinear approximation is appropriate. R_{s0} is the sum of the deviations of the chain monomers in each coordinate dimension from the center of the square or cube in that dimension. For a perfect square with $a = b = \nu^{1/2}$ monomers per side,

$$\begin{aligned} R_{s0}^2 &= (4/a) \sum_{k=1}^{(a-1)/2} k^2 \\ &= (1/6)(a^2 - 1)(a + 1)/a \\ &\approx \nu/6. \end{aligned}$$

Hence, if $\nu \gg 1$, $R_{s0}^2 \approx \nu/6$. By the same logic, in 3D $R_{s0}^2 \approx \nu^{2/3}/4$. R_{l0}^2 in 2D and 3D are approximated by the same method, substituting n for ν .

Rosenbluth weighting is used to correct for the bias introduced when the MC simulation efficiency is increased by choosing each step only from the unoccupied neighboring sites (Rosenbluth & Rosenbluth, 1955). The Rosenbluth weight of a lattice configuration is accumulated as the product of factors determined at each step. The factor at each step is the ratio of the number of adjacent unoccupied sites relative to the maximum possible number of adjacent unoccupied sites. Meirovitch (1982) has noted that Rosenbluth weighting necessarily overestimates the number of configurations that can be completed, therefore each completed configuration is overweighted. We note further that the overweighting errors increase as the global constraints on the chain increase: the weighting error is greater in 2D than in 3D, greater for the SCM than the LCM, and greater for compact chains than for open chains. This suggests that the estimates for excess entropy reported here, due to addition of side chains and packing density, should be taken as limiting minimum values.

The distributions of contacts in Figures 7 and 10 were determined by the MC sampling of chains of length $n = 50$ main-chain monomers. We note that all $\langle \Phi_I \rangle$ are essentially independent of chain length for n greater than ~ 20 . Further, the dependencies of $\langle \Phi_I \rangle$ on ρ_s are nearly identical for chain lengths n greater than ~ 20 in 3D and ~ 30 in 2D.

We compare SCM configurations of length $n = 40$ rather than $n = 50$ with protein structures from the PDB in order to increase the size of the maximally compact sample population accessible with our algorithm. For the PDB data, each main-chain monomer position is defined by the C_α coordinates, and each side-chain monomer position is defined by the C_β coordinates. Contact weights in the SCM are the product of the Rosenbluth weight of the given configuration and the inverse sixth power of the distance between the contacting monomers. Contact weights in the PDB are determined by the distance calculation alone. Because contact weights are determined from distances between pairs of carbon coordinates and the C_α 's and C_β 's of amino acids successive in the primary sequence are always close in space, the lowest order contacts included in these calculations are contact $o = 2$. Because the lattice produces an oscillation between contacts of odd and even contact orders, the contact frequencies are averaged over contact order intervals of 2 bonds.

Table 2. Lattice constants^a

	2D (SE)	3D (SE)
	$\Omega_L(n-1) = c_L \sigma_L^{(n-1)}$ $\Omega_S(n-1) = \Omega_L(n-1) c_S \sigma_S^{(n-2)}$	
c_L	1.837 (0.008)	1.362 (0.004)
σ_L	2.737 (0.001)	4.808 (0.001)
c_S	1.230 (0.006)	0.366 (0.009)
σ_S	1.501 (0.001)	3.546 (0.001)
	LCM: $\langle R_S^2 \rangle = a_L(n-1)^{b_L}$ SCM: $\langle R_S^2 \rangle = a_S(2n-3)^{b_S}$	
a_L	0.135 (0.005)	0.182 (0.004)
b_L	1.447 (0.004)	1.178 (0.003)
a_S	0.074 (0.015)	0.104 (0.017)
b_S	1.462 (0.009)	1.198 (0.010)

^a Ω_L and Ω_S are determined in exhaustive simulations for $n = 4-16$ in 2D and $n = 4-10$ in 3D. LCM and SCM $\langle R_S^2 \rangle$'s are determined from exhaustive simulations for $n = 13-16$ in 2D and $n = 9, 10$ in 3D and from MC sampling for $n = 20, 30, 40$, and 50 in 2D and 3D. c_L , σ_L , c_S , σ_S , a_L , b_L , a_S , and b_S are determined by the linear least-squares fit to the logarithmic equations. Symmetry conventions have been chosen to represent the characteristics of proteins. Because proteins have polarity, the termini of the chain are considered to be distinguishable. To account for side-chain degrees of freedom, configurations related by reflection are considered to be distinguishable, as are configurations related by rotation in the planes of the lattice. The following operations relate configuration counts herein to previous work from this laboratory. In Chan and Dill (1989a, 1989b), $N = n - 1$ = the number of bonds, and total reflections and rotations in any of the planes are not counted as distinct configurations. Therefore, $4(2\Omega(N) - 1) = \Omega_I$ for $N = n - 1$. In Lau and Dill (1989), n = the number of monomers, but total reflections and rotations are not counted as distinct configurations. Therefore, $4(2\Omega(N) - 1) = \Omega_I(n)$. In Chan and Dill (1989a), the symmetry conventions are the same as in this work, but the numbers of conformations are given as a function of the number of bonds, N .

Acknowledgments

We thank D.O.V. Alonso, H.S. Chan, K.M. Fiebig, D.P. Yee, and K. Yue for their extraordinary generosity and helpful discussions; L. DeYoung, P. Harkins, O.B. Ptitsyn, G.D. Rose, D. Shortle, and T.C. Terwilliger for valuable comments; W.A. Lim and F.M. Richards for their useful insights and for sharing their manuscript prior to publication; and the NIH for financial support.

References

- Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Seivers R, eds. *Crystallographic databases—Information content, software systems, scientific applications*. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107–132.
- Baker EN, Hubbard RE. 1984. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol* 44:97–179.
- Baldwin EP, Hajiseyediavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodations of variants that repack the core of T4 lysozyme. *Science* 262:1715–1718.
- Behe MJ, Lattman EE, Rose GD. 1991. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? *Proc Natl Acad Sci USA* 88:4195–4199.
- Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
- Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.
- Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science* 247:1306–1310.
- Bruch MD, Dhingra MM, Gierasch LM. 1991. Side chain-backbone hydrogen bonding contributes to helix stability in peptides derived from an alpha-helical region of carboxypeptidase A. *Proteins Struct Funct Genet* 10:130–139.
- Camacho CJ, Thirumalai D. 1993. Minimum energy compact structures of random sequences of heteropolymers. *Phys Rev Lett* 71:2505–2508.
- Casassa EF, Berry GC. 1966. Angular distribution of intensity of Rayleigh scattering from comblike branched molecules. *J Polym Sci Polym Phys Ed* 4:881–897.
- Chan HS, Dill KA. 1989a. Compact polymers. *Macromolecules* 22:4559–4573.
- Chan HS, Dill KA. 1989b. Intrachain loops in polymers: Effects of excluded volume. *J Chem Phys* 90:492–509.
- Chan HS, Dill KA. 1990a. Origins of structure in globular proteins. *Proc Natl Acad Sci USA* 87:6388–6392.
- Chan HS, Dill KA. 1990b. The effects of internal constraints on the configurations of chain molecules. *J Chem Phys* 92:3118–3135.
- Chan HS, Dill KA. 1991. “Sequence space soup” of proteins and copolymers. *J Chem Phys* 95:3775–3787.
- Chan HS, Dill KA. 1994. Transition states and folding dynamics of proteins and heteropolymers. *J Chem Phys*. Forthcoming.
- Chang IS, Shapir Y. 1988. Collapse transition of branched polymers with a tunable number of loops. *Phys Rev B* 38:6736–6740.
- Creamer TP, Rose GD. 1992. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci USA* 89:5937–5941.
- Daoud M, Joanny JF. 1981. Conformation of branched polymers. *J Physique* 42:1359–1371.
- deGennes PG. 1979. *Scaling concepts in polymer physics*. Ithaca, New York: Cornell University Press.
- Derrida B, Herrmann HJ. 1983. Collapse of branched polymers. *J Physique* 44:1365–1376.
- Dickman R, Schieve WC. 1984. Lattice animal specific heats and the collapse of branched polymers. *J Physique* 45:1727–1730.
- Dill KA. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24:1501–1509.
- Dill KA. 1990. Dominant forces in protein folding. *Biochemistry* 29:7133–7155.
- Domb C. 1963. Excluded volume effects of 2 and 3 dimensional lattice models. *J Chem Phys* 38:2957.
- Eden D, Matthew JB, Rosa JJ, Richards FM. 1982. Increase in apparent compressibility of cytochrome c upon oxidation. *Proc Natl Acad Sci USA* 79:815–819.
- Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* 255:178–183.
- Finney JL. 1975. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol* 96:721–732.
- Flesia S, Gaunt DS, Soteros CE, Whittington SG. 1992. General model for collapse in lattice animals. *J Phys A: Math Gen* 25:L1169–L1172.
- Flory PJ. 1953. *Principles of polymer chemistry*. Ithaca, New York: Cornell University Press.
- Freed KF. 1987. *Renormalization group theory of macromolecules*. New York: Wiley.
- Gaunt DS. 1991. Contact models of a collapsing branched polymer. *Physica A* 177:146–152.
- Gavish B, Gratton E, Hardy CJ. 1983. Adiabatic compressibility of globular proteins. *Proc Natl Acad Sci USA* 80:750–754.
- Gupta SK, Forsman WC. 1973. Probability distributions for the radii of gyration of short, branched, random-flight chains. *Macromolecules* 6:285–287.
- Gurd FR, Rothgeb TM. 1979. Motions in proteins. *Adv Protein Chem* 33:73–165.
- Handel TM, Williams SA, Degradó WF. 1993. Metal ion dependent modulation of the dynamics of a designed protein. *Science* 261:879–885.
- Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262:1680–1685.
- Karlin S, Blaisdell BE, Bucher P. 1992. Quantile distributions of amino acid usage in protein classes. *Protein Eng* 5:729–738.
- Karplus M, McCammon JA. 1981. The internal dynamics of globular proteins. *CRC Crit Rev Biochem* 9:293–349.
- Karplus M, Shakhnovich EI. 1992. Protein folding: Theoretical studies of thermodynamics and dynamics. In: Creighton TE, ed. *Protein folding*. New York: W.H. Freeman and Company. pp 127–195.
- Keefe LJ, Sondek J, Shortle D, Lattman EE. 1993. The alpha aneurism: A structural motif revealed in an insertion mutant of staphylococcal nuclease. *Proc Natl Acad Sci USA* 90:3275–3279.
- Kim KS, Fuchs JA, Woodward CK. 1993. Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry* 32:9600–9608.
- Klapper MH. 1971. On the nature of the protein interior. *Biochim Biophys Acta* 229:557–566.
- Kundrot CE, Richards FM. 1987. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. *J Mol Biol* 193:157–170.
- Lam PM. 1988. Scaling analysis of the collapse of branched polymers. *Phys Rev B* 38:2813–2815.
- Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of protein. *Macromolecules* 22:3986–3997.
- Lau KF, Dill KA. 1990. Theory for protein mutability and biogenesis. *Proc Natl Acad Sci USA* 87:638–642.
- Lesk AM. 1991. *Protein architecture: A practical approach. The practical approach series*. New York: IRL Press.
- Lim WA, Sauer RT. 1989. Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature* 339:31–36.
- Lipson JEG. 1991. A Monte Carlo simulation study on long-chain combs. *Macromolecules* 24:1327–1333.
- Lyu PC, Wemmer D, Zhou HX, Pinker RJ, Kallenbach NR. 1993. Capping interactions in isolated α -helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. *Biochemistry* 32:421–425.
- Madras N, Soteros CE, Whittington SG, Martin JL, Sykes MF, Flesia S, Gaunt DS. 1990. The free energy of a collapsing branched polymer. *J Phys A: Math Gen* 23:5327–5350.
- Matthews BW. 1987. Genetic and structural analysis of the protein stability problem. *Biochemistry* 26:6885–6888.
- McCammon JA, Harvey SC. 1987. *Dynamics of proteins and nucleic acids*. New York: Cambridge University Press.
- McCrackin FL, Mazur J. 1981. Configuration properties of comb-branched polymers. *Macromolecules* 14:1214–1220.
- McRee DE, Redford SM, Getzoff ED, Lepock JR, Hallewell RA, Tainer JA. 1990. Changes in crystallographic structure and thermostability of a Cu, Zn superoxide dismutase mutant resulting from the removal of a buried cysteine. *J Biol Chem* 265:14234–14241.
- Meirovitch H. 1982. A new method for simulation of real chains: Scanning future steps. *J Phys A: Math Gen* 15:L735–L741.
- Miller R, Danko CA, Fasolka MJ, Balazs AC, Chan HS, Dill KA. 1992. Folding kinetics of proteins and copolymers. *J Chem Phys* 96:768–780.

- Miranker A, Radford SE, Karplus M, Dobson CM. 1991. Demonstration by NMR of folding domains in lysozyme. *Nature* 349:633–636.
- Nemirovsky AM, Bawendi MG, Freed KF. 1987. Lattice models of polymer solutions: Monomers occupying several lattice sites. *J Chem Phys* 87:7272–7284.
- Novotny J, Brucoleri R, Karplus M. 1984. An analysis of incorrectly folded protein models: Implications for structure predictions. *J Mol Biol* 177:787–818.
- Pickett SD, Sternberg MJ. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol* 231:825–839.
- Presta LG, Rose GD. 1988. Helix signals in proteins. *Science* 240:1632–1641.
- Ptitsyn OB. 1987. Protein folding: Hypothesis and experiments. *J Protein Chem* 6:273–293.
- Ptitsyn OB. 1992. The molten globule state. In: Creighton TE, ed. *Protein folding*. New York: W.H. Freeman and Company. pp 243–300.
- Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI. 1990. Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett* 262:20–24.
- Radford SE, Dobson CM, Evans PA. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* 358:302–307.
- Rey A, Skolnick J. 1993. Computer modeling and folding of four-helix bundles. *Proteins Struct Funct Genet* 16:8–18.
- Richards FM. 1974. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J Mol Biol* 82:1–14.
- Richards FM. 1977. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng* 6:151–176.
- Richards FM, Lim W. 1994. An analysis of packing in the protein folding problem. *Q Rev Biophys*. Forthcoming.
- Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of α helices. *Science* 240:1648–1652.
- Roder H, Elove GA, Englander SW. 1988. Structural characterization of folding intermediates in cytochrome *c* by H-exchange labelling and proton NMR. *Nature* 335:700–704.
- Roovers J. 1979. Synthesis and dilute solution characterization of comb polystyrenes. *Polymer* 20:843–849.
- Rose GD, Wolfenden R. 1993. Hydrogen bonding, hydrophobicity, packing and protein folding. *Annu Rev Biophys Biomol Struct* 22:381–415.
- Rosenbluth MN, Rosenbluth AW. 1955. Monte Carlo calculation of the average extension of molecular chains. *J Chem Phys* 23:356–359.
- Shakhnovich EI, Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a 1st-order phase transition. *Biopolymers* 28:1667–1680.
- Shortle D, Chan HS, Dill KA. 1992. Modeling the effects of mutations on the denatured states of proteins. *Protein Sci* 1:201–215.
- Singh J, Thornton JM. 1990. SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. *J Mol Biol* 211:595–615.
- Singh J, Thornton JM. 1992. *Atlas of protein side-chain interactions*. New York: Oxford University Press.
- Solc K. 1973. Statistical mechanics of random-flight chains. IV. Size and shape parameters of cyclic, star-like and comb-like chains. *Macromolecules* 6:378–385.
- Stickle DF, Presta LG, Dill KA, Rose GD. 1992. Hydrogen bonding in globular proteins. *J Mol Biol* 226:1143–1159.
- Varadarajan R, Richards FM. 1992. Crystallographic structures of ribonuclease S variants with nonpolar substitution at position 13: Packing and cavities. *Biochemistry* 31:12315–12327.
- Wagner G. 1983. Characterization of the distribution of internal motions in the basic pancreatic trypsin inhibitor using a large number of internal NMR probes. *Q Rev Biophys* 16:1–57.
- Wilkinson MK. 1986. Branched polymers: Exact enumeration study of three-dimensional lattice animals classified by valence distribution. *J Phys A: Math Gen* 19:3431–3441.
- Wüthrich K, Wagner G. 1978. Internal motions in globular proteins. *Trends Biochem Sci* 3:227–230.
- Yee DP, Dill KA. 1993. Families and the structural relatedness among globular proteins. *Protein Sci* 2:884–899.
- Ziman JM. 1979. *Models of disorder: The theoretical physics of homogeneously disordered systems*. New York: Cambridge University Press.
- Zimm BH, Stockmayer WH. 1949. The dimensions of chain molecules containing branches and rings. *J Chem Phys* 17:1301–1314.