J. Mol. Biol. (1994) **241**, 557–573

Does Compactness Induce Secondary Structure in Proteins?
A Study of Poly-alanine Chains Computed by Distance Geometry

David P. Yee[1], Hue Sun Chan[1], Timothy F. Havel[2] and Ken A. Dill[1]

[1]*Department of Pharmaceutical Chemistry*
University of California, San Francisco, CA 94143-1204, U.S.A.

[2]*Department of Biological Chemistry and Molecular Pharmacology*
Harvard Medical School, 240 Longwood Avenue
Boston, MA 02115, U.S.A.

A few years ago, lattice model studies indicated that compactness could induce polymer chains to develop protein-like secondary structures. Subsequent off-lattice studies have found the amounts of induced structure to be relatively small. Here we use distance geometry to generate random conformations of compact poly-alanine chains of various chain lengths. The poly-alanine chains are subjected only to compactness and excluded volume constraints; no other energies or conformational propensities are included in the chain generation procedure. We find that compactness leads to considerable stabilization of secondary structure, but the absolute amount of secondary structure depends strongly on the criteria used to define helices and sheets. By loose criteria, much secondary structure arises from compactness, but by strict criteria, little does. The stabilization free energy of secondary structure provided by compactness, however, appears to be independent of criteria. Since real helices and sheets in proteins can be identified by strict criteria, we introduced small energy perturbations to compact poly-alanine chains using the AMBER force field. Small refinements produced good α-helices. For β-sheets, however, larger refinements are necessary. Compactness appears to impart stability, but not much structural specificity, to secondary structures in proteins. Compactness acts more like diffusion as a force, a result of ensemble statistics, than like pair interactions such as hydrogen bonding.

Keywords: secondary structure; compactness; distance geometry; protein folding

1. Introduction

Nearly 50% of the residues in globular proteins are in either α-helix or β-sheet (Levitt & Greer, 1977; Kabsch & Sander, 1983). What are the forces that stabilize secondary structures in globular proteins? Helices and sheets can be identified on the basis of their hydrogen bonding patterns (Kabsch & Sander, 1983), so it is reasonable to expect that hydrogen bonds play some role. But since the short helices that are predominant in globular proteins are not very stable when isolated in solution, other aspects of the surrounding protein must help stabilize them (Dill, 1990). Furthermore, recent results on the refolding of cytochrome *c* suggest that the protein population refolds at the same rate as molecular collapse (Sosnick *et al.*, 1994). This result precludes the formation of substantial amounts of stable secondary structure prior to collapse. A few years ago, it was proposed that compactness in single polymer chains could contribute substantially to the formation of regular internal structure in proteins (Chan & Dill, 1989, 1990*a*,b). Based on exhaustive simulations of short chains on two-dimensional square lattices and systematic conformational searches on three-dimensional cubic lattices, it was found that the amount of secondary structure increases sharply with the compactness of a polymer chain. Those studies found roughly the same chain length distributions of helices and sheets in compact lattice chain conformations as in the protein structures in the Protein Data Bank (PDB†; Abola *et al.*, 1987).

Subsequently, other studies of lattice models (Kolinski & Skolnick, 1992) and off-lattice models (Gregoret & Cohen, 1991; Hao *et al.*, 1992; Socci *et al.*, 1994; Hunt *et al.*, 1994) have explored in greater detail the role of compactness in inducing secondary structure. They are summarized briefly below.

Gregoret & Cohen (1991) generated native-like random chains using an off-lattice rotational

† PDB, Protein Data Bank.

isomeric model of proteins. Chains were modeled as linear strings of residues. Each residue excluded a spherical volume. Random chain conformations were constructed using a backtrace Monte Carlo procedure in which virtual bond and torsion angles were selected from the distribution observed in real protein structures. Chains were restricted to lie within ellipsoids defined by a formula derived from real proteins. Gregoret & Cohen observed an increase in secondary structure with increasing compactness, but only found significant amounts of secondary structure, by their criteria, when the chains are 30% more compact than real proteins.

Hao *et al.* (1992) also used a Monte Carlo backtrace procedure to generate random chains. They too used a simplified representation of proteins where side-chains were represented as spheres. The chains have random amino acid sequences. Compact chain conformations were generated with torsion angles selected either randomly or weighted to reflect local interactions imposed by the peptide bond. Hao *et al.* compared a bond vector correlation function for real proteins with the random chains and concluded that the observed bond vector correlations in real proteins can be reproduced best when (1) the chains are constrained to be compact, and (2) intra-residue interactions are included.

Kolinski & Skolnick (1992) used a high-resolution lattice model of poly-alanine and poly-valine chains. Local and non-local alanine–alanine and valine–valine interactions were derived from angular correlations of side-group vectors in proteins taken from the PDB. Hydrogen bonding was included in the form of two terms: (1) an attractive hydrogen bond term, and (2) a cooperative interaction for adjacent hydrogen bonds. Conformational space was explored by a Monte Carlo dynamics algorithm. The poly-alanine chains readily formed helical structures at low temperatures and exhibited a cooperative helix-coil transition. Similar results were obtained when (1) local bond correlations and hydrogen bonding terms were used, (2) when a non-local hydrophobic term was added, and (3) when the hydrophobic and hydrogen bonding terms were used without local bond correlations. In no case did the poly-alanine chains collapse into compact states.

The poly-valine chains, however, collapsed into compact, β-sheet-like conformations when using bond correlation, hydrogen bonding and hydrophobic interactions. When only the hydrogen bonding and hydrophobic terms were used in the absence of local bond correlations, they collapsed to compact states consisting of about 25% helix and no β-sheet. When only the hydrophobic interaction was used, the chains collapsed into compact states, but no appreciable secondary structure was found. They concluded that collapse alone was not sufficient to produce secondary structure.

Socci *et al.* (1994) developed a simplified model in which each residue was represented by a single point. Random conformations were generated by minimizing a potential that included a covalent term for chain connectivity, an r^{12} term for excluded volume, and a radius of gyration term to drive compactness. Two constants in the potential that control (1) the balance between covalent and non-covalent forces, and (2) strength of excluded volume, were derived by examining real proteins. They looked for repeating structure by defining a function that measured correlations between dihedral angles between different points along the protein chain. They could not detect any well-defined repeating patterns that resembled secondary structure if the chains were only constrained to adopt compact conformations.

Hunt *et al.* (1994) extended the earlier work of Gregoret & Cohen (1991). They used an all-atom, off-lattice model of protein structure. Conformational space was searched using a Monte Carlo simulated annealing method in which the ϕ/ψ angles of a randomly selected residue were reassigned from a distribution of ϕ/ψ angles derived from the PDB. Simplified energy functions were used that included: (1) a radius of gyration term to induce compactness in the chain, (2) an energy term for hydrogen bond effects, or (3) a combination of (1) and (2). Hunt *et al.* observed an increase in secondary structure when only the compactness term was used. When a combination of the compactness and hydrogen bond terms were used, they were able to generate highly compact conformations with amounts of secondary structure comparable to real proteins.

Why is there a need for yet another study? While these studies have contributed considerable insight into the role of packing in secondary structure formation, they have also raised new questions. Although they represent proteins more accurately than the original lattice model, they, too, are simplified models. In order to avoid simplified side-chain and lattice models, we study here an all-heavy-atom model of poly-alanine, of chain lengths 50, 100 and 150, in a continuum representation. Chirality is taken into account, whereas it is not in some of the earlier studies. The use of backtrace Monte Carlo and constraining ellipsoids can introduce conformational bias, so here we use distance geometry to constrain the conformations. Our statistics indicate that the conformations generated by distance geometry are relatively unbiased. In lattice models, compactness and secondary structure can be defined with little ambiguity, but in off-lattice models it is not so simple. What conformations should be called helices and sheets? Does the amount of secondary structure observed depend on the criteria used to define it? Here, we explore these issues.

2. Methods

(a) *Generating unbiased compact poly-alanine conformations*

Poly-L-alanine conformations were generated using the DG-II distance geometry program (Havel, 1991), with sequential tetrangle inequality bound smoothing, randomized metrization using a uniform distribution, and embedding in 4 dimensions followed by 10,000 steps of

dynamical simulated annealing refinement in which the superfluous dimension was eliminated. The resulting convergence rate averaged about 80%, and the annealing was simply repeated (using differential initial velocities) on non-convergent conformations until they converged. In addition to the constraints needed to obtain the proper covalent geometry and chirality, a uniform upper bound was imposed on all the non-bonded distances, thereby effectively packing the polypeptide chain into a sphere whose radius is half of that upper bound. Previous extensive studies of the randomized metrization method on unconstrained poly-L-alanine chains (Havel, 1990) have demonstrated that it yields coordinates that can readily be refined to polypeptide chains whose statistical properties are in accord with the statistical mechanics of chain molecules. It is therefore reasonable to expect that the compact, self-avoiding polypeptide chains generated in this paper, for which few theoretical predictions can be made, are likewise satisfactorily unbiased.

(b) *Determining compactness*

Determining compactness is simple for lattice models, but is more difficult for more realistic models. For example, Gregoret & Cohen (1991) calculated the radius, R, of a hypothetical ideal spherical protein based on average properties of amino acids and proteins using the expression:

$$R = \left[\frac{3 \times N \times m}{4\pi \times \rho_P \times 10^6 \times N_A}\right]^{1/3} \times 10^{10}, \quad (1)$$

where N is the number of residues, m is the average molecular mass of an amino acid (110 g/mol), ρ_P is the average density of globular proteins (1·4 g/mol), N_A is Avogadro's number, and 10^6 ml/m^3 and 10^{10} Å/m are unit conversion factors. Compactness was then varied by scaling the chain volume using the equation:

$$V = \varepsilon \times \frac{4\pi abc}{3}, \quad (2)$$

where a, b and c are the principal axes of the ellipsoid and $abc = R^3$. Gregoret & Cohen suppose the condition $\varepsilon = 1$ represents real native proteins, and $\varepsilon < 1$ represents higher densities.

However, using their ideal values in equation (1), the volume occupied by a single amino acid is 130·4 Å^3. But in their chain generation procedure, a residue is allowed to be placed as close as 4·25 Å from any other non-bonded residue. This implies that the volume excluded by one chain segment is at most only $4\pi/3 \times (4{\cdot}25/2)^3 = 40{\cdot}2$ Å^3. Hence, excluded volume is considerably underestimated, by $130{\cdot}4 - 40{\cdot}2 = 90{\cdot}2$ Å^3 per amino acid. Whereas they estimate a native-like radius of gyration for their 64-mers of 11·0 Å, if we use their excluded volume of 40·2 Å^3 per residue, then the radius that contains all the residues of the maximally compact state would be:

$$R = \left(\frac{3(40{\cdot}2\ \text{Å}^3)(64)}{4\pi(0{\cdot}74)}\right)^{1/3} = 9{\cdot}4\ \text{Å}, \quad (3)$$

which is equivalent to a radius of gyration of $R_G = \sqrt{3/5}R = 7{\cdot}3$ Å (Chan & Dill, 1991). The factor of 0·74 in the denominator of equation (3) approximates the maximal packing density for the model system: it is both the theoretical maximum packing density of close-packed spheres of the same size and the packing density observed in crystals of small organic molecules (Richards, 1974). Note that if 130·4 Å^3 is used in equation (3) instead of 40·2 Å^3 for the volume excluded per amino acid, we obtain $R_G = 10{\cdot}8$ Å $\approx 11{\cdot}0$ Å. While their studies show less than 20% secondary structure with constraining radii of 11·0 Å, their studies with radii of gyration of 7·3 Å show more than 35% secondary structure.

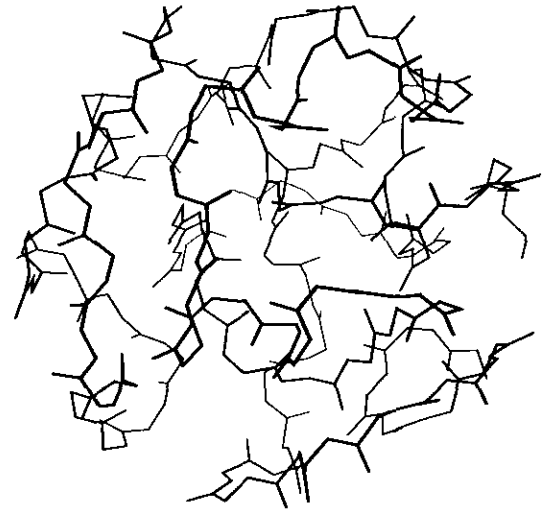


Figure 1. Conformation of a near maximally compact poly-alanine 100-mer generated by the distance geometry method.

In our study, chains were constrained to be enclosed by spheres of specific radii by the distance geometry procedure. Specifying a large radius yields relatively open chains, while specifying a small radius gives compact chains. We generated several sets of random chains varying in both chain length and compactness. Table 1 lists the chain lengths, the constraining radii, and the average radius of gyration for each set of poly-alanine conformations. Figure 1 shows an example of a near maximally compact poly-alanine 100-mer.

By 2 different criteria, the most compact sets of these chains are nearly maximally compact. First, we use the criterion of Chothia (1975), by which the mean volume of alanine, when buried in the protein core, is 91·5 Å^3. The minimal radius of a maximally compact chain is given by:

$$\tfrac{4}{3}\pi R^3 = N \times 91{\cdot}5\ \text{Å}^3, \quad (4)$$

where R is the radius in ångstrom units and N is the chain length. Solving equation (4) for $N = 50$, 100 and 150 gives

Table 1
Poly-alanine chains

No. of residues	Constraining radius (Å)	$\langle R_G \rangle$† (Å)	No. of conformations
50	10·0	7·96	100
	11·0	8·62	100
	12·5	9·60	100
	15·0	11·14	100
100	12·5	10·11	100
	14·0	11·12	100
	16·5	12·72	50
	20·0	14·96	50
150	15·0	11·92	25
	17·0	13·26	25
	20·0	15·25	25

† R_G is based on C^α atom positions.

maximally compact radii of 10·3, 13 and 14·8 Å, respectively. The most compact sets of 50, 100 and 150-mers have constraining radii of 10, 12·5 and 15 Å, so by this criterion, the chains are nearly maximally compact.

Second, we use the criterion of Maiorov & Crippen (1992) who derived an expression for the minimal radius of gyration for polypeptide chains:

$$R_{\min} = -1{\cdot}26 + 2{\cdot}79 \times N^{1/3}. \quad (5)$$

According to equation (5), the radius of gyration of a maximally compact 50, 100 and 150-mer should be 9·0, 11·7 and 13·6 Å^3. Thus, by this second criterion too, the poly-alanine chains are nearly maximally compact.

(c) *Defining helices and sheets*

There is no single correct way to determine secondary structure. There are several published methods for identifying secondary structures in proteins (Levitt & Greer, 1977; Kabsch & Sander, 1983; Richards & Kundrot, 1988). While they are largely consistent with one another, and differ mainly at the ends of helices and sheets, their differences, nevertheless, can be considerable (Dev, 1987). In the present study, we use 3 very different methods for assigning secondary structures. We use Define (Richards & Kundrot, 1988), based on inter-residue distances, DSSP (Kabsch & Sander, 1983) based on hydrogen bonding patterns, and a topological contact (TC) method (Chan & Dill, 1990*a*,*b*) based on patterns of inter-residue contacts. We are able to use DSSP effectively only in cases when we include hydrogen bonding interactions in the AMBER refinements, since it is not applicable to models without defined hydrogen bonds. The methods are described below.

(i) *Define*

Define (Richards & Kundrot, 1988) identifies secondary structure by comparing the inter-residue distance map of a chain segment with the inter-residue distance maps of an ideal α-helix and extended strand. If the difference distance map between the chain segment and the ideal helix or strand is below some threshold, then the residues in the segment are identified as participating in secondary structure. Error thresholds and mismatch limits were set to their default values. Since this procedure does not match strands to locate or identify sheets, Define is used only to identify helices and strands.

(ii) *DSSP*

DSSP (Kabsch & Sander, 1983) is the most stringent definition we used. It is based primarily on hydrogen bonding patterns. Helices, for example, are defined in terms of repeats of hydrogen bonds between residues separated by 3, 4 or 5 residues. Sheets are identified by locating hydrogen bonds between residues that are not close in sequence.

(iii) *Topological contact (TC)*

One definition of secondary structure that does not depend on geometries or bond angles, but depends only on the topology of neighboring contacts was given by Chan & Dill (1990*a*,*b*). It was implemented and applied to model and real protein structures by Gregoret & Cohen (1991), who showed that it correctly identifies helices and sheets in proteins. By identifying secondary structures on the basis of specific contacts, the TC method is similar to the definition of secondary structure used by Levitt & Greer (1977).

In this method, a helix is identified by patterns of specific contacts between residues. A contact between 2 residues is defined when the distance between the C$^\alpha$ atom positions is less than some cutoff value. In this work, we primarily use a cutoff of 5·5 Å. A helix is defined if either of the following conditions holds:

(1) $[(i, i+3), (i+2, i+5)]$ and
(2) $[(i, i+3), (i, i+5), (i+1, i+6), (i+2, i+7), (i+4, i+7)]$.

Since (2) is mainly used to identify helices on lattices, definition (1) was the main identifier of helices in the poly-alanine chains. Antiparallel sheets are defined by contacts between residues $[(i, j+2), (i+1, j+1), (i+2, j)]$. Parallel sheets are defined by contacts between residues $[(i, j), (i+1, j+1), (i+2, j+2)]$. All residues involved in a putative sheet must be in an extended conformation specified by certain lower bounds on the interior virtual bond angles. To avoid double counting, all helical residues must have interior virtual bond angles less than these bounds. We primarily use a bond angle cutoff of 100°.

(d) *Energy minimization*

In some of the studies described below, the constrained conformations were further refined using energy minimization. The AMBER (Weiner *et al.*, 1984, 1986) potential as implemented in the Insight II 2.2.0/Discover 2.9 molecular modeling package from Biosym Technologies was used in all energy minimizations. 1–4 non-bonded interactions were scaled by 0·5.

(e) *Structural comparison and clustering*

To determine whether energy minimizations substantially perturbed the conformations, we used CONGENEAL (Yee & Dill, 1993), a computer algorithm that calculates the structural dissimilarity between conformations. CONGENEAL represents chains in terms of their weighted distance maps. The weighted distance map of a protein chain that has N residues is an $N \times N$ matrix in which each matrix element (i, j) is a weight, w, equal to the distance, $d_{i,j}$, between the α-carbon atoms of residues i and j, raised to a power, -2 (i.e. $w_{i,j} = d_{i,j}^{-2}$). The essential feature of the weighting function is that residues that are close together in space are given more weight than residues that are distant in space.

Comparisons are performed by superimposing 2 weighted distance maps and summing the absolute differences between corresponding distance weights. The dissimilarity is defined as the sum of the absolute differences between corresponding inter-residue distance weights normalized by the average of the summed distance weights for each of the 2 distance maps. So for 2 chain conformations R and S, the dissimilarity between the chains is given by:

$$d(R, S) = \frac{\sum\sum |r_{ij}^{-2} - s_{ij}^{-2}|}{\frac{1}{2}\left(\sum\sum r_{ij}^{-2} + \sum\sum s_{ij}^{-2}\right)}. \quad (6)$$

For P structures, all $P(P-1)/2$ possible pairs of structures are first compared, then clustered using a hierarchical clustering method. Clustering is performed by iteratively grouping P structures into larger and larger groups. The similarity between groups is defined to be the average of the similarities between the members of one group compared to the members of another. At the beginning of the clustering process, the 2 most similar structures are merged into a single group. Then the next most similar pair of structures or groups of structures are merged. This process continues until all structures are finally merged into a single group at the top of the tree.

The final tree-like structure shows the inter-relatedness of the structures within the set.

3. Results

Our first question was whether the distance geometry method for generating compact conformations led to local biases in the bond angles. We found that it did not. While the studies of Gregoret & Cohen (1991) and Hunt *et al.* (1994) focus mainly on chains with intrinsic peptide bond angles, taken from the distribution observed in the PDB, our interest here was to explore a slightly different question of whether polymers without local bond biases could be induced to have secondary structures. Intrinsic peptide bond propensities must surely help in the formation of secondary structures, but we are also interested to know whether other types of polymers might be induced to form secondary structures. Gregoret & Cohen (1991) constructed conformations by adding C^{α} positions to a growing chain. The position of each new C^{α} atom was selected by picking virtual bond angles (α) and virtual torsion angles (τ), which reproduced the α/τ distribution derived from a database of real protein structures. If, after a given number of attempts, a new C^{α} position could not be assigned, the previously placed C^{α} atom was reassigned. Figure 2(A) shows the distribution of α/τ angles reproduced by the chain generation method of Gregoret & Cohen. Their bond conformations are concentrated in specific regions of α/τ space. In contrast, Figure 2(B) shows that the α/τ distributions obtained from our constrained poly-alanine chains are uniform, reflecting the absence of local interactions in the chain generation procedure.

Because our poly-alanine chains are based on an all-heavy-atom representation, our ϕ/ψ plots are consistent with ϕ/ψ (Ramachandran) plots of hard sphere models, and include chirality, in contrast to earlier lattice (Chan & Dill, 1990*a*,*b*) and off-lattice (Gregoret & Cohen, 1991) models, for which the chain representations are too simple to include chirality.

(a) *ϕ/ψ dihedral angle distributions*

In proteins, ϕ/ψ angles are clustered into two primary regions: an α_R region and a β region. These areas are associated with α-helix and β-sheet, respectively. In addition, the α_L region is somewhat populated in real proteins. Figure 3(A) shows a ϕ/ψ map derived from proteins in the PDB. The α_R and β regions are prominent on the left-hand side of the Figure. The less populated cluster on the upper right-hand side of the Figure represents the α_L region of the map.

Since the observed distribution of dihedral angles in real proteins is generally consistent with simple steric avoidance (Ramachandran & Sasisekharan, 1968), it is not unexpected that the dihedral angle distributions of the poly-alanine chains generated in this work are similar to the dihedral angle distribution observed in real proteins. For example, there is a tilted oval near the center of the ϕ/ψ plot (i.e. $(\phi, \psi) \approx (0, 0)$) that is not populated (see Figure 3(B)). This region represents steric clashes between the oxygen atom of residue i with either (1) the carbonyl carbon of atom $i+1$, or (2) with the amide hydrogen of residue $i+1$. The blank region centered at $\phi = 0$ and extending from $\psi = -180$ to $+180$ is due to contacts between the oxygen atoms of residues i and $i+1$. There is a forbidden region for all ψ centered at $\phi \approx 120$, which is due to interactions of the peptide backbone with side-chain

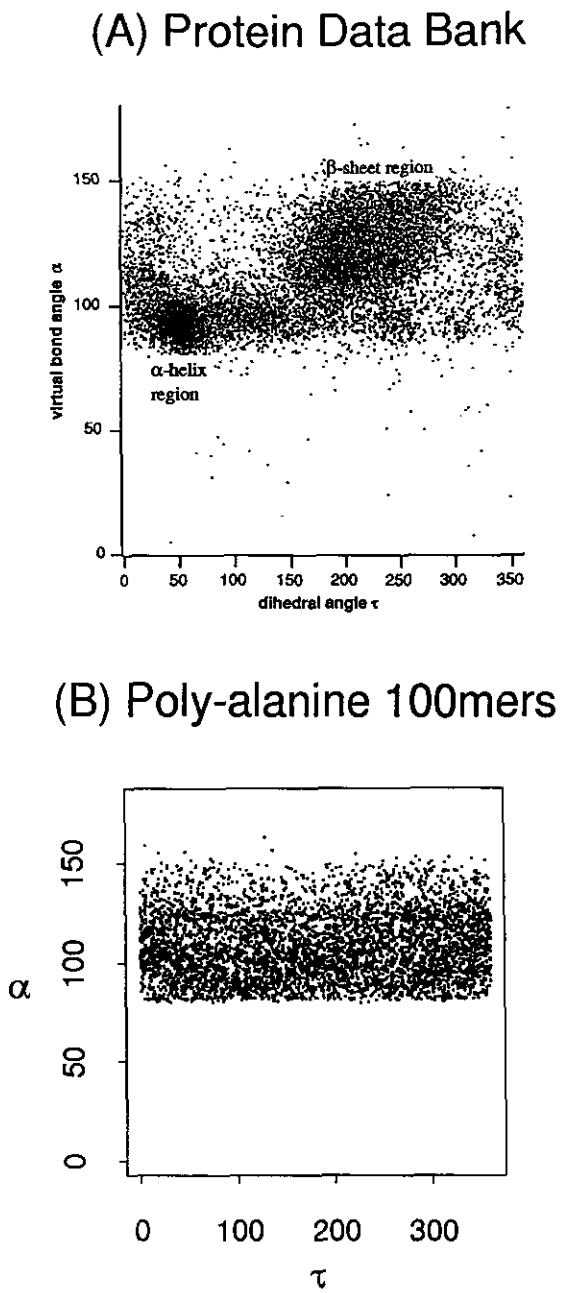


Figure 2. Distributions of virtual bond angle α *versus* virtual torsion angle τ. (A) Distribution derived from real protein structures and used by Gregoret & Cohen (1991). (B) Distribution derived from random, compact 100-mer poly-alanine conformations.

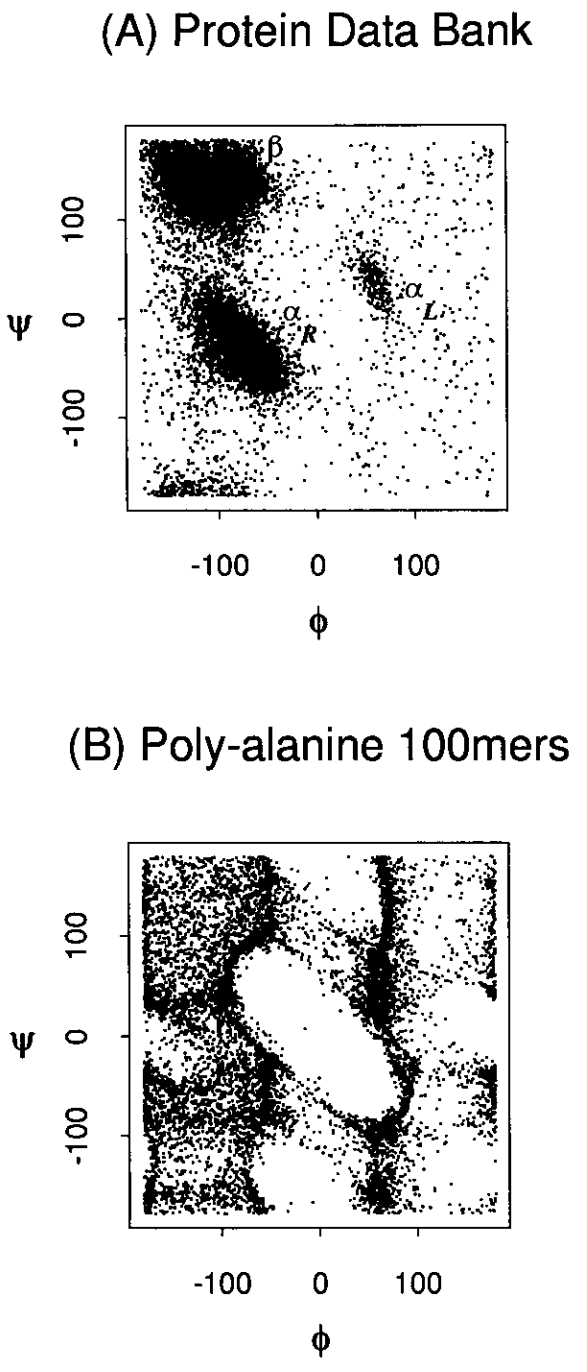


Figure 3. (A) Distribution of ϕ/ψ angles derived from the crystal structures obtained from the Brookhaven Protein Data Bank. (B) ϕ/ψ distribution of 100 random, near maximally compact 100-mer poly-alanine chains. The ϕ/ψ distributions of all other sets of poly-alanine chains (both compact and open) are nearly identical.

atoms. Comparison of Figure 3(A) and (B) shows that most poly-alanine ϕ/ψ angles fall loosely within the β, α_R and α_L regions of the ϕ/ψ map.

There are also differences, however, between the ϕ/ψ distribution of the poly-alanine chains and the distribution from real proteins. First, the distribution of ϕ and ψ angles is more diffuse and spread out than in real proteins. Second, whereas proteins have a small α_L region corresponding to a left-handed helix, these poly-alanine chains have a long continuous strip of populated dihedral angles defined by $(\phi, \psi) \approx (60-80, -180-180)$. Some of the clustering on the right-hand side of the ϕ/ψ map is due to the nature of the error functions used in the distance geometry procedure. This is because the basin of attraction leading into the allowed region is large relative to the size of that region (Havel, 1990). Nonetheless, the detailed distribution of dihedral angles observed in real proteins cannot be explained completely on the basis of excluded volume effects. Therefore, although the ϕ/ψ angle combinations observed in the poly-alanine chains do not involve steric violations, some of the ϕ/ψ angle combinations are not energetically favorable conformations. This result is consistent with molecular dynamics simulations of model alanyl dipeptides which show that, in solution, the regions on the left-hand side of the ϕ/ψ map (in particular, the α_R and β regions) are strongly favored relative to regions on the right-hand side of the map (Anderson & Hermans, 1988).

We found that the dihedral angle distributions in poly-alanine are independent of chain length and compactness. Thus, chain compactness appears not to influence local conformational preferences at the level of single pairs of dihedral angles.

(b) *Secondary structure in poly-alanine chains*

How much secondary structure is there in the confined poly-alanine chains? We used three different criteria to identify helices and sheets. According to DSSP, the maximally compact chains do not contain regular secondary structures in the form of α-helices or β-sheets. This result is not unexpected since DSSP relies on the identification of hydrogen bonds to locate secondary structure. Since no energetics were used to generate these poly-alanine chain conformations, there are few residue-to-residue orientations which can be identified as being hydrogen bonded. Applying Define to the maximally compact chains shows that the chains have significant strand content. About 14% of residues in all the poly-alanine chains (compact and non-compact) are in extended strand conformations. By the TC criterion, the compact chains have significant antiparallel sheet content. Both Define and TC methods agree that there is very little α-helix content. Figure 4 summarizes the absolute amount of secondary structure in poly-alanine 50, 100 and 150-mers determined by a variety of criteria.

By all criteria, the amount of secondary structure increases with chain compactness. Since there is no *a priori* correct distance to use as a cutoff for defining a contact for the TC criterion, we systematically varied: (1) the cutoff distances defining a contact, and (2) the bond angle constraint required for defining an extended strand conformation. Reducing the cutoff distance defines helices more stringently. Consistent with the earlier lattice

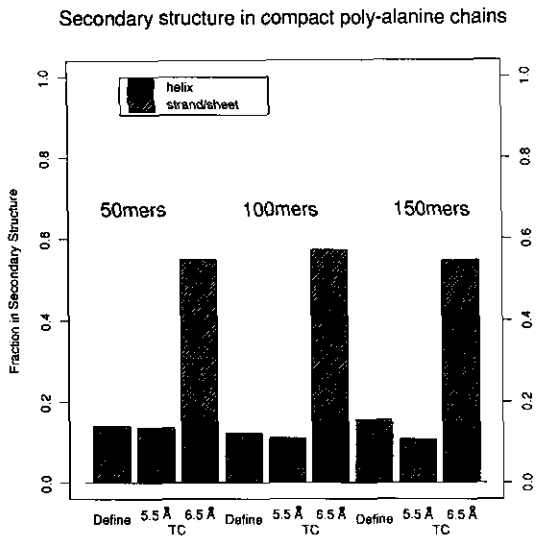


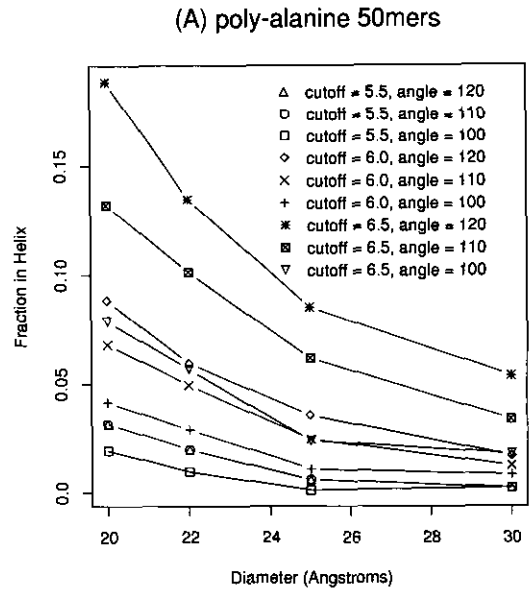
Figure 4. The absolute amount of secondary structure in compact poly-alanine 100-mers. Define identifies the amount of extended strand and helix. TC identifies sheet and helix. The result of using both a 5·5 Å and 6·5 Å cutoff with the TC definition of secondary structure is shown.

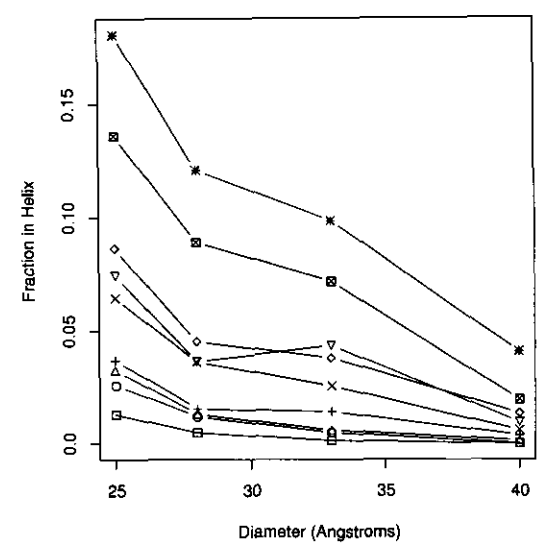
studies (Chan & Dill, 1990*a*,*b*), Figures 5 and 6 show that the amount of helix and sheet defined by these various criteria increases with compactness. The Figures also show the effect of altering the stringency of the secondary structure definitions. By varying the cutoff parameter from 5·5 to 6·5 Å, and the bond angle of strands from 100 to 120°, the amount of observed secondary structure can vary over a large range: from about 1% to 20% for helix and from 3% to 50% for sheet. For all criteria, the helix and sheet content increases with compactness.

While the amount of *structure* is strongly dependent on the defining criterion, the amount of *stabilization free energy* is not. Figure 7 shows that there is a free energy stabilizing secondary structures that comes from compactness. The loss in free energy, $-\Delta G_{\text{compactness}}$, is the logarithm of the ratio of the fraction of secondary structure residues in the compact state to that of the open state:

$$-\Delta G_{\text{compactness}}/kT = \ln\left[\frac{(\textit{fraction in } 2^\circ \textit{ structure})_{\text{compact}}}{(\textit{fraction in } 2^\circ \textit{ structure})_{\text{open}}}\right]. \quad (7)$$

Figure 7 shows the amount of secondary structure as a function of compactness plotted on a logarithmic scale. The y-axis corresponds to free energy lost in units of kT. Surprisingly, despite the wide range in the absolute numbers of residues in





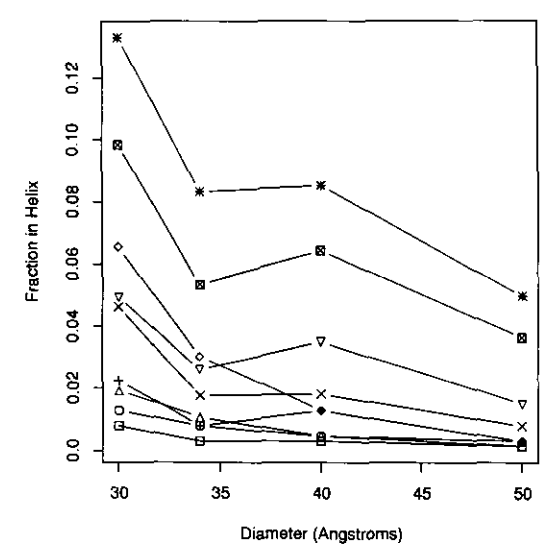
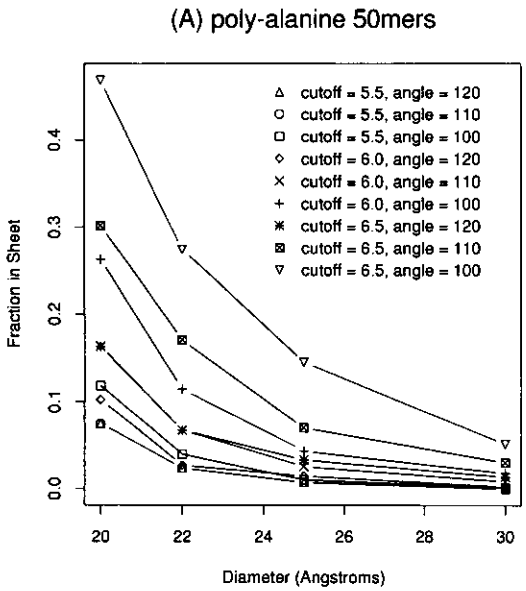
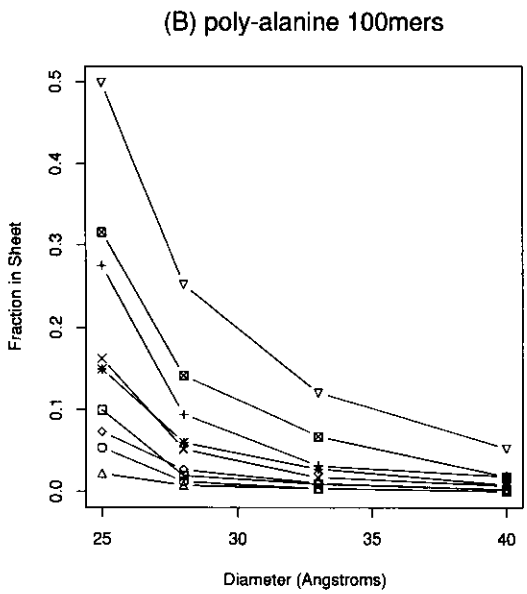
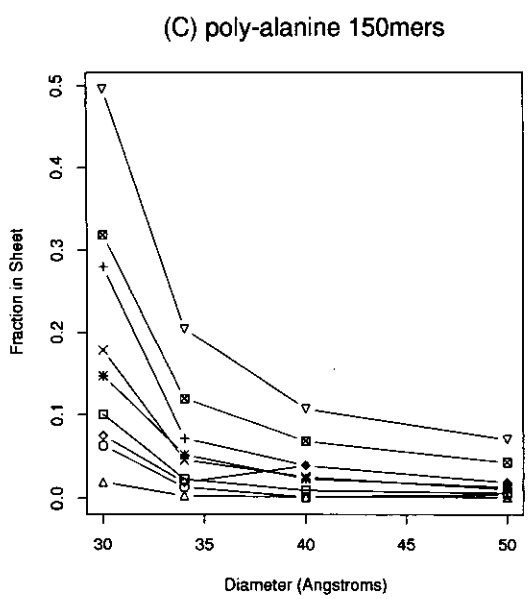


Figure 5. Amount of helix as a function of compactness as determined by the TC definition. Cutoffs defining a contact were varied from 5·5 to 6·5 Å. The maximum bond angle allowed for residues to be assigned to a helix conformation was varied from 100° to 120°. (A) Poly-alanine 50-mers; (B) poly-alanine 100-mers; (C) poly-alanine 150-mers.







secondary structure by different criteria, the change in free energy with compactness appears to be largely independent of the criteria used.

More informative than the total amount of secondary structure is its distribution. Figure 8(A) to (I) shows the distributions of compact (blue) and open (pink) conformations as a function of the number of residues in secondary structure. Each panel represents a different set of criteria used with the TC definition of secondary structure. From the upper left to the lower right represents decreasing stringency of criterion. In all cases, it is clear that the absolute amount of secondary structure is higher in the compact conformations than in the open conformations. In addition, for poly-alanine chains in compact ensembles, there are more conformations with large numbers of residues in secondary structure than with small numbers. The reverse is true for ensembles of open chain conformations. Almost no open chains have large amounts of secondary structure.

From these data we can estimate the free energy of stabilization of secondary structures by compactness. We define a structural unit as six residues, since this is the minimal requirement of the TC criterion for defining a helix or sheet. The calculation is performed by dividing the fraction of compact conformations with N to $N+5$ residues in secondary structure by the fraction of open conformations with N to $N+5$ residues in secondary structure. The logarithm of this ratio yields an enhancement factor in units of kT. Note that when there are no conformations with N to $N+5$ residues in secondary structure, an enhancement factor is undefinable. Hence, our statistics are limited since the sets of compact poly-alanine chains represent only a small sampling of all possible compact states. Nonetheless, the enhancement factors for poly-alanine 50-mers as a function of the number of secondary structural units are shown in Figure 9: each *additional* 6-mer structural unit of secondary structure is stabilized (favored) by compactness by about 2 kT. This estimate is largely independent of the identifying criteria used.

(c) *Energy minimization*

So far we have described all-heavy-atom poly-alanine chains that are only under the influence of the compactness constraint imposed by a constraining radius. The only constraints are that (1) no two atoms may occupy the same space (i.e. excluded volume), and (2) the entire chain conformation must be configured within the bounds imposed by a constraining radius. No other biases or

Figure 6. Amount of sheet as a function of compactness as determined by the TC definition. Cutoffs defining a contact were varied from 5·5 to 6·5 Å. The minimum bond angle required for residues to be assigned to a sheet conformation was varied from 100° to 120°. (A) Poly-alanine 50-mers; (B) poly-alanine 100-mers; (C) poly-alanine 150-mers.

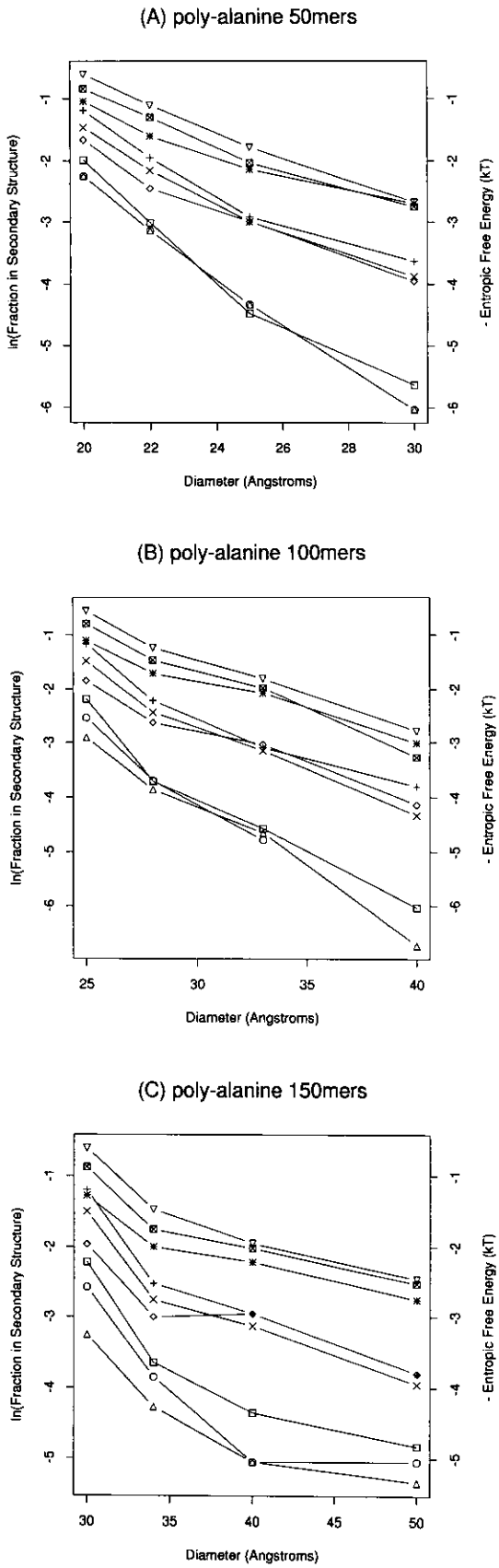


Figure 7. Amount of total secondary structure (helix + sheet) as a function of compactness as in Figures 5 and 6, except that the y-axis is plotted on a logarithmic scale. (A) Poly-alanine 50-mers; (B) poly-alanine 100-mers; (C) poly-alanine 150-mers.

energies have been included. The results described above show that strict criteria do not identify very much structure in these compact poly-alanine chains. The helices and sheets in these constrained poly-alanine chains do not look very protein-like. That is, compactness is not a force that specifically drives the formation of α-helices and β-sheets, but rather it stabilizes broad classes of conformations that include helices and sheets as subsets. Compactness acts to favor certain topological repeats, such as $(i, i+3)$ contacts, but a considerable range of bond angles and geometries are consistent with this. The stabilization of secondary structures afforded by compactness can be viewed in the way that diffusion can be viewed as a driving force. It is not a specific pair interaction; it is a global property of an ensemble. This driving force is not structurally specific, like hydrogen bonds or other pair interactions are. We and others (Hunt *et al.*, 1994; Honig *et al.*, 1993) believe that both types of interaction, the stabilization afforded by compactness, and the structural specificity afforded by hydrogen bonding and local propensities, are required to lock in structures as specific as the α-helices and β-sheets observed in real proteins. Compactness appears to give stability, but not conformational specificity, to secondary structures in globular proteins.

Our results show that compactness increases the amount of ordered structure in compact polymers, but how far are the conformations from energy minima of more realistic secondary structures? The distribution of ϕ/ψ angles in the compact poly-alanine chains deviate substantially from the distribution of dihedral angles observed in real proteins. The degree to which a chain can move is severely limited, since the excluded volume effect is very strong for the compact chains. We now ask whether a small perturbation of random compact poly-alanine conformations by the AMBER force field is sufficient to induce the poly-alanine structures to look more like proteins.

The poly-alanine chains that were generated by distance geometry were subsequently energy minimized using the AMBER potential. In addition to using the unmodified AMBER potential, the force field was modified in several ways to see how specific restraints would alter the ϕ/ψ map of the minimized poly-alanine chains. We tested four types of energy minimization: (1) the unmodified AMBER potential; (2) increasing the strength of the hydrogen bond; (3) adding a torsion angle force toward ϕ/ψ minima observed in alanine dipeptide simulations; and (4) adding constraints to favor the formation of α-helix-like hydrogen bonds. The results of these minimizations are presented below.

(1) Figure 10(A) and (B) shows the ϕ/ψ maps for sets of ten compact poly-alanine chains before and after 10,000 cycles of conjugate gradient minimization using the unmodified AMBER potential. The minimization procedure makes the ϕ/ψ distribution more protein-like in several ways. The ϕ/ψ angles in the α_R region becomes slightly more concentrated. The strip of torsion angles between $\phi \approx 60-80$

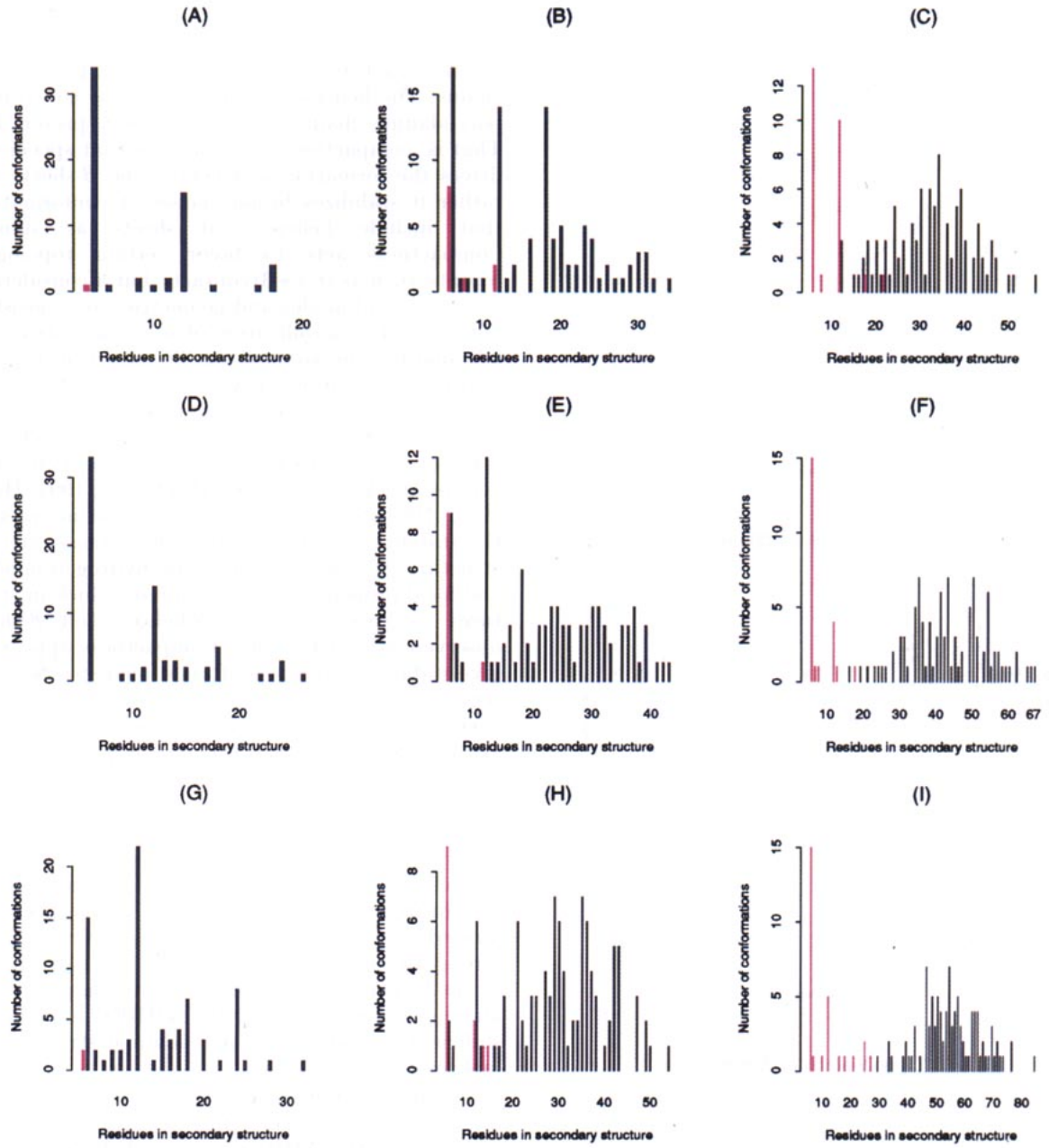


Figure 8. Histograms showing the number of poly-alanine 100-mer conformations as a function of the number of residues in secondary structure. The blue bars correspond to compact poly-alanine chains (25 Å sphere). The pink bars correspond to open poly-alanine chains (40 Å sphere). Each panel represents a different criteria set used with the TC definition of secondary structure. The criteria used are as follows: (A) contact cutoff, 5·5 Å, bond angle cutoff, 120°; (B) contact cutoff, 5·5 Å, bond angle cutoff, 110°; (C) contact cutoff, 5·5 Å, bond angle cutoff, 100°; (D) contact cutoff, 6·0 Å, bond angle cutoff, 120°; (E) contact cutoff, 6·0 Å, bond angle cutoff, 110°; (F) contact cutoff, 6·0 Å, bond angle cutoff, 100°; (G) contact cutoff, 6·5 Å, bond angle cutoff, 120°; (H) contact cutoff, 6·5 Å, bond angle cutoff, 110°; (I) contact cutoff, 6·5 Å, bond angle cutoff, 100°.

becomes less continuous. In general, the overall distribution of torsion angles is less diffuse and more focused into distinct minima. Nevertheless, differences remain between the energy minimized ϕ/ψ distribution shown in Figure 10(B) and the real protein distribution. Since any energy minimization strategy seeks the *nearest* energy minimum, the refinement found only the closest ϕ/ψ combinations which would lower the overall energy of the system. For example, the map shows an increased concentration of (ϕ,ψ) angles near $(-70,60)$ and $(65,-60)$, which correspond to the formation of hydrogen bonds between $O_{i-1}-HN_{i+1}$. These regions have been identified as energy minima in alanine dipeptide model studies in vacuum and in solution (Pettitt & Karplus, 1988; Head-Gordon *et al.*, 1991; Tobias & Brooks, 1992). In addition, the concentration of dihedral angles in the α_L region $(\phi,\psi \approx 55,45)$ is increased.

(2) We increased the strength of the hydrogen

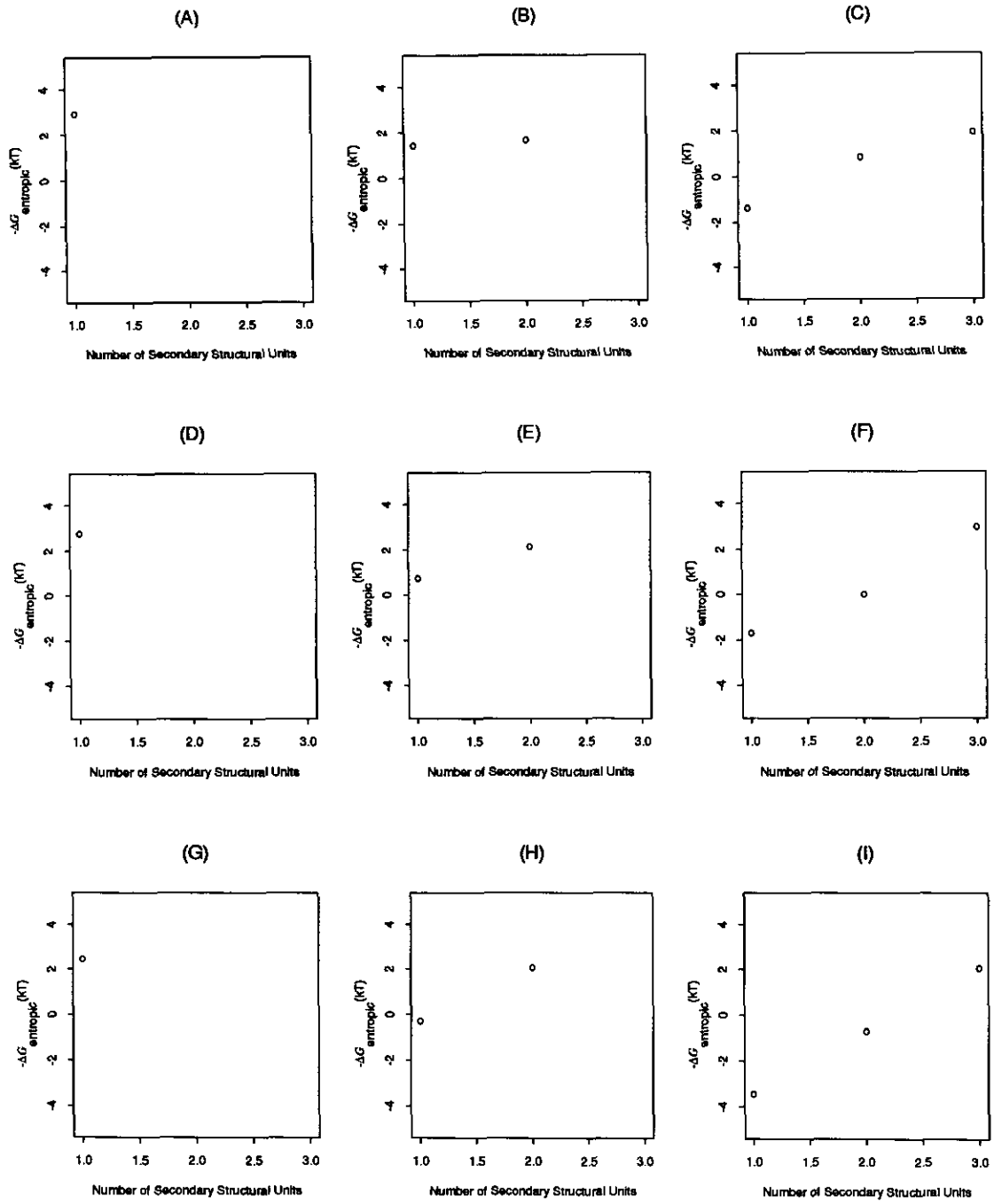


Figure 9. Enhancement factors for poly-alanine 50-mer conformations transferred from an open ensemble to a compact one as a function of 6 residue secondary structural units (i.e. 1 implies 6 residues in secondary structure, 2 implies 7 to 12 residues, etc.). Each panel represents a different criteria set which is given in the legend to Figure 8.

bond by factors of 1·50, 2·00 and 3·00. In each case, the ϕ/ψ distributions were more diffuse (for example, see Figure 10(C)). There is a significant increase in the population of the β-sheet region of the ϕ/ψ map. Some ϕ/ψ angles, however, populate regions of the map that are not sampled by real proteins and several of the distances between adjacent C^{α} atoms become closer than the 3·8 Å expected for *trans*-peptide bonds. Hence, increasing the strength of the hydrogen bond beyond its physical value can result in rather severe distortion of the standard peptide bond geometry.

(3) Next, we introduced a torsion angle forcing potential. Molecular dynamics simulations of alanine dipeptides have mapped out ϕ and ψ angles that correspond to local energy minima (Anderson & Hermans, 1988). They indicate that two primary free energy minima exist at $(\phi, \psi) \approx (-110, 120)$ and $(-120, -40)$, which correspond to the β region and α_R region, respectively. Note that the "helix" ϕ minimum is offset by 50° (to −120) from the ideal value for an α-helix. In the present work, we added forcing potentials in the form of an extra harmonic energy term to shift the dihedral angles ϕ and ψ to

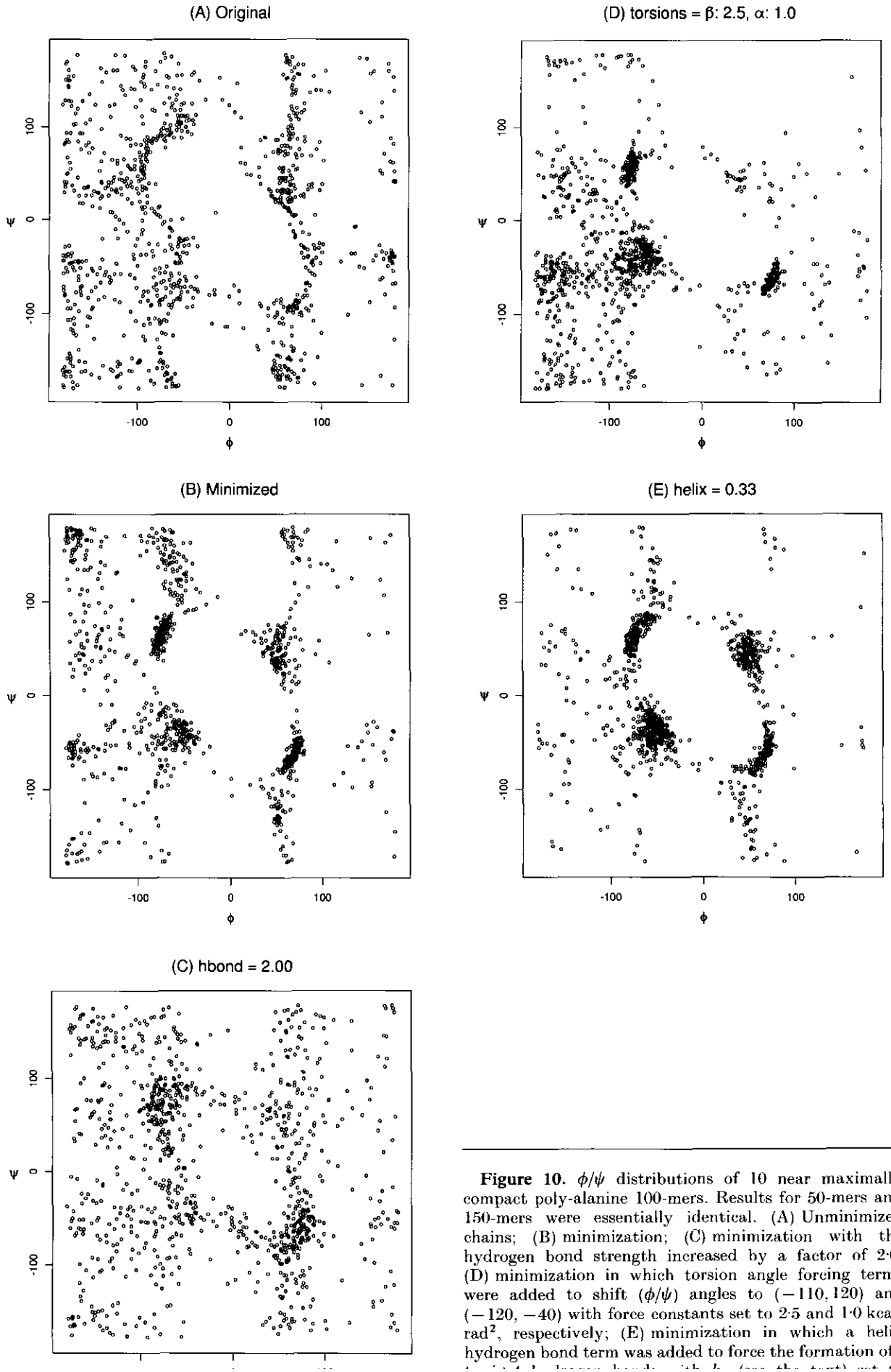


Figure 10. ϕ/ψ distributions of 10 near maximally compact poly-alanine 100-mers. Results for 50-mers and 150-mers were essentially identical. (A) Unminimized chains; (B) minimization; (C) minimization with the hydrogen bond strength increased by a factor of 2·0; (D) minimization in which torsion angle forcing terms were added to shift (ϕ/ψ) angles to $(-110, 120)$ and $(-120, -40)$ with force constants set to 2·5 and 1·0 kcal/rad^2, respectively; (E) minimization in which a helix hydrogen bond term was added to force the formation of i

these values. The magnitude of the forcing potential ranges from 1·0 to 5·0 kcal/rad^2 for the β region and from 0·4 to 2·0 kcal/rad^2 for the α_R region.

Figure 10(D) shows an example of energy minimizing poly-alanine chains subject to the torsion angle forcing term. There was a noticeable shift of torsion angles toward the α_R region of the ϕ/ψ map. Even though the force constants favoring the β region were larger than for the α_R region, there was little enhancement of ϕ/ψ angles in the extended β region.

(4) Finally, a term was added to force conformations toward the formation of hydrogen bonds between O_i and HN_{i+4}. The extra energy term was a flat-bottomed, skewed biharmonic function as implemented within Discover version 2.9 by Biosym Technologies. The form of the term is given by:

$$E_{\text{constrain}}(r) = \begin{cases} E_{L,\max} + f_{\max}(r_{L,\max} - r) & r < r_{L,\max} \\ k_L(r - r_L)^2 & r_{L,\max} < r < r_L \\ 0 & r_L < r < r_U \\ k_U(r - r_U)^2 & r_U < r < r_{U,\max} \\ E_{U,\max} + f_{\max}(r - r_{U,\max}) & r_{U,\max} < r. \end{cases} \quad (8)$$

Here, the value of r_L was 1·7 Å. k_L was fixed at 5·0 kcal/mol Å^2 to prevent the atoms from getting too close to one another. r_U was taken to be 2·5 Å. Values of k_U ranged from 0·10 to 2 kcal/mol Å^2. $r_{L,\max}$ and $r_{U,\max}$ were selected such that:

$$f_{\max} = 10{\cdot}0 \text{ kcal mol}^{-1} \text{ Å}^{-2} = 2k_L(r_L - r_{L,\max}) = 2k_U(r_{U,\max} - r_U). \quad (9)$$

$E_{L,\max}$ and $E_{U,\max}$ were given by:

$$\text{and} \quad E_{L,\max} = k_L(r_{L,\max} - r_L)^2 \quad E_{U,\max} = k_U(r_{U,\max} - r_U)^2. \quad (10)$$

Figure 10(E) shows that there is an increase in the population of dihedral angles in the α_R and α_L regions, even for very small force constants. The results indicate that both the α_R and α_L regions of the ϕ/ψ map are consistent with the formation of helical hydrogen bonds. The observation of more ϕ/ψ angles in the α_L region of the poly-alanine chains relative to real proteins is due to the fact that the starting (unminimized) poly-alanine conformations have ϕ/ψ angles that populate sterically allowable, but energetically unfavorable, regions on the right-hand side of the ϕ/ψ map. When subjected to the force field, the ϕ/ψ angles on the right-hand side of the map move toward the closest minimum (i.e. α_L).

(d) *Secondary structure in energy minimized poly-alanine chains*

Figure 11 shows how introducing the energy perturbations affects the amount of secondary structure in the confined poly-alanine chains. Minimization alone using the unmodified AMBER potential slightly enhances the amount of secondary structure in the poly-alanine chains. According to Define, the strand content is lower in the minimized structures. Strengthening local hydrogen bonds reduces the chain extension. On the other hand, according to both DSSP and TC, energy minimization increases the sheet content. The TC criterion also detected an increase in helix content. Increasing the strength of the hydrogen bond slightly increases the helix content according to both DSSP and TC. The TC criterion also detects a large increase in sheet content. Part of the increased sheet content was due to severe distortions as the poly-alanine chains become very compact from the strong hydrogen bonds.

By all secondary structure criteria, the amount of α-helix increases after adding either the torsion angle term or the α-helical hydrogen bond force, even when the force constants are small (e.g. 0·1 kcal/mol Å^2). As discussed above, when the helical hydrogen bond force is added, many ϕ/ψ angles populate the α_L region of the ϕ/ψ map. This means that left-handed helices can form and may be identified as α-helix by secondary structure detection methods that are not sensitive to handedness such as Define and TC. DSSP, however, is sensitive to handedness and parallels the results obtained from Define and TC. Thus, sterically constrained compact poly-alanine conformations require only small perturbations to reach α-helices. (We show below that these perturbations are small.)

It is much more difficult, however, to force the formation of β-sheets. Simple energy minimization may not introduce sufficient perturbation to induce sheets. Sheets require coordination of different strands to come together, whereas helices appear to require only local readjustments that occur readily with small energetic perturbations.

(e) *Structural similarity of energy minimized chains*

Here we show that energy minimization does not cause large perturbations of the poly-alanine chain conformations. We compared pairwise all conformers of five poly-alanine chains that were energy minimized using the strategies described above. For each chain, there are 11 conformational variations: the original plus the result of ten different energy minimization runs. Structural similarity was evaluated using CONGENEAL (Yee & Dill, 1993). The pairwise comparison data were then used to construct a relatedness tree using hierarchical clustering. Figure 12 shows that after minimization, each poly-alanine chain falls within its own "family" of structures. Even when the forcing potentials are very large, the perturbed structures will cluster with their original poly-alanine parent. This indicates that after energy minimization, each poly-alanine chain retains enough of its original chain fold to be classified within its proper family. Also by molecular graphics we observed that the placements of turns in these structures are relatively unchanged. Even with a large helical force, for example, the chains never straighten out into a long helix.

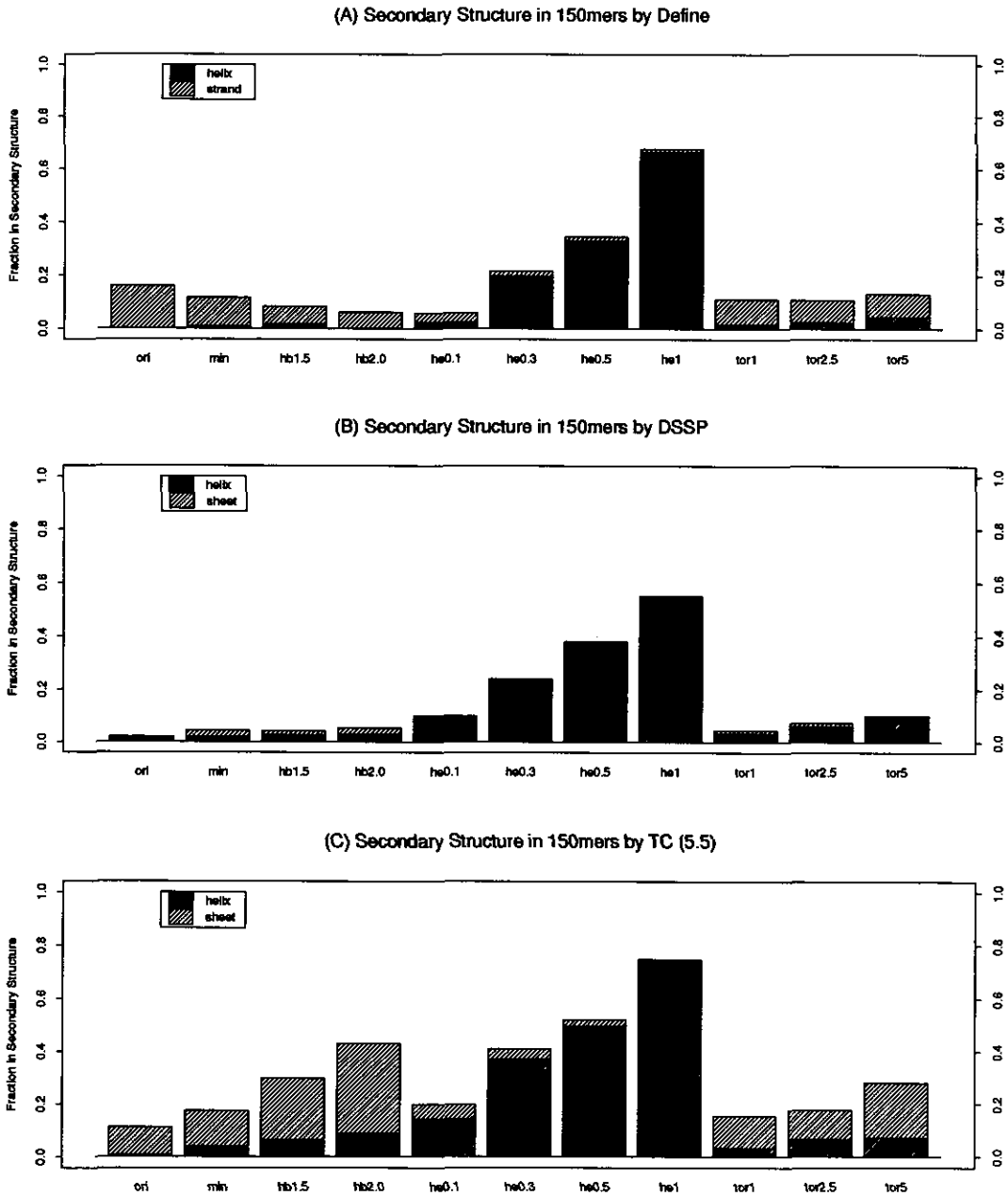


Figure 11. Histogram showing absolute amounts of secondary structure in maximally compact poly-alanine 150-mers after various minimization strategies. (A) Secondary structure determined by Define; (B) secondary structure determined by DSSP; (C) secondary structure determined by TC with 5·5 Å, 100° cutoff. ori, original conformation; min, AMBER minimization; hb1.5, hydrogen bond strength scaled by 1·5; hb2.0, hydrogen bonds scaled by 2·0; he0.1, helix term with force constant, k_H, set to 0·1 kcal/mol Å^2; he0.3, $k_H = 0{\cdot}33$; he0.5, $k_H = 0{\cdot}5$; he1, $k_H = 1{\cdot}0$; tor1, torsion angles terms for β and α regions set to 1·0 and 0·4 kcal/rad^2; tor2.5, torsion terms with force constants equal to 2·5 and 1·0; tor5, torsion terms with force constants equal to 5·0 and 2·0.

When open chains were energy minimized using the strategies described above, the resulting conformations were only distantly related to the original starting conformation. That is, energy perturbations on open, relatively unconstrained chains led to large changes in the overall fold of the original structure.

(f) *Comparison with other studies*

Despite the differences in methodology and model, this study is in general agreement with the studies of Gregoret & Cohen, Hao *et al.*, Socci *et al.* and Hunt *et al.* in several main conclusions. First, it shows that packing is not structurally specific: there is considerable conformational diversity. By strict criteria, only a small fraction of the stabilized secondary structures are recognizable α-helices and β-sheets. Compactness cannot account for why α-helices are more prevalent than 3_{10} helices, but it can account for why helices in general are more stable in compact conformations than in open conformations. Second, we find that only small energetic perturbations from the confined poly-alanine

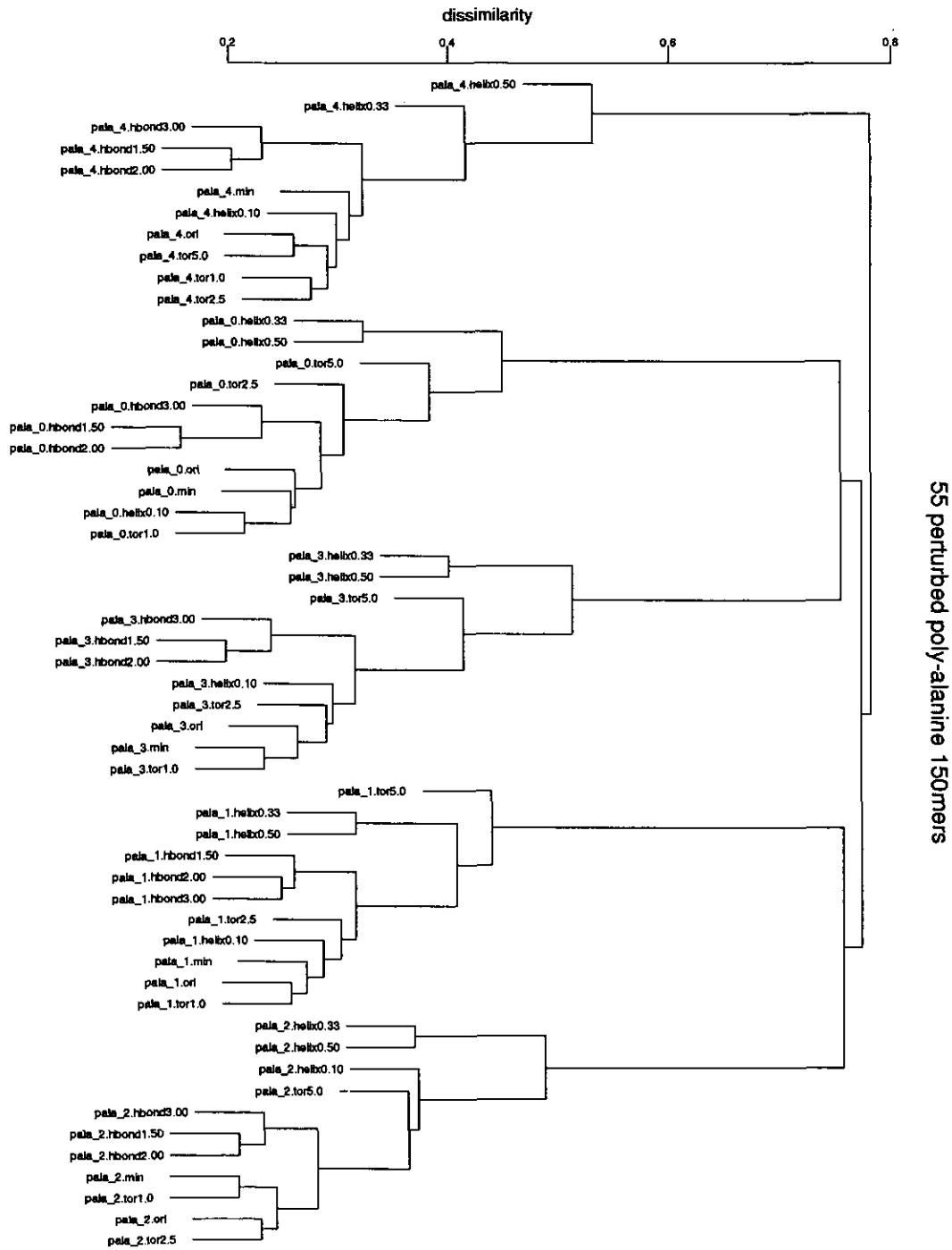


Figure 12. Tree showing inter-relatedness of energy minimized poly-alanine 150-mers. Each poly-alanine chain clusters with its "parent" structure even when forcing potentials are large.

chains are required to give good α-helical conformations, although larger perturbations appear to be required to get good β-sheets. Third, consistent with the earlier lattice studies of Chan & Dill (1989, 1990*a*,*b*), most of these studies show that compactness enhances secondary structure, and hence it provides a driving force. However, Kolinski & Skolnick (1992) appear to disagree with this view. They say: "Thus . . . the secondary structure seen in the folded state is predominantly the result of short-range interactions, or conformational propensities, which are more or less in accord with packing requirements in the dense globular state". They make two arguments. First, they note that the distribution of distances between residues i and $i+3$ in poly-valine chains look very similar for both compact and random coil ensembles when only the hydrophobic interaction is used. In addition, their distributions also resembled the distribution at low temperatures when only local (i.e. $|i-j|<7$) interactions were included. Observing no differences in these distributions, they concluded that secondary structure does not come from compactness. Second, their poly-valine chains collapsed into compact

β-sheet-like globular states when local conformational preferences, hydrogen bonds and non-local interactions were included. When local conformational preferences were left out, the poly-valine chains collapsed into compact states which were highly helical. When only the non-local interaction is used, the poly-valine chains adopted compact, "disordered" conformations. Since they used a strict criterion of secondary structure based on hydrogen bonding patterns, it is not surprising that they did not see any (hydrogen-bond-based) secondary structure when only non-local interactions were used. This is consistent with our present results. That Kolinski & Skolnick could generate more secondary structures in compact states using more complex potentials does not prove that compactness is not important. In our view, compactness in their study effectively eliminates the large number of non-compact conformations. The compact conformations that remain have a larger proportion of residues which can participate in secondary structure. In order to test the hypothesis that compactness induces secondary structure within the context of their model, Kolinski & Skolnick might have asked how much local and hydrogen bond driving forces are necessary to get native-like secondary structures in the presence and absence of compactness. We believe compactness will reduce the driving force necessary to achieve native-like structures.

Finally, we disagree with a recent suggestion by Karplus & Shakhnovich (1992) that the hypothesis of compactness enhancement of secondary structure (Chan & Dill, 1990*a*,*b*) contradicts Flory's theorem (Karplus & Shakhnovich, 1992). Flory's theorem postulates that the spatial distribution of monomers of an individual polymer chain in a dense multiple-chain polymer melt would resemble that of a random flight (Flory, 1949; de Gennes, 1979). The theorem does not address the conformations of single compact chains. In addition, the Flory theorem addresses only distributions of monomer pairs, and not the multiple monomer correlations in secondary structures. Hence, the current results are not inconsistent with the Flory theorem.

4. Conclusions

We have modeled proteins as all-heavy-atom poly-alanine chains of lengths 50, 100 and 150 monomers. They have been configured randomly, subject only to: (1) confinement to different radii of gyration by distance geometry; and (2) steric constraints. No local propensities, energies, or other biases have been explicitly included.

We find that compactness enhances secondary structure by several different criteria of identifying helices and sheets. The total amount of secondary structure observed depends strongly on the criteria used and can vary over a range from nearly zero to 50% for the same chain conformation. Our results agree with other studies that have used strict criteria, in that the secondary structures induced from compactness alone are neither protein-like nor as narrow a class of structures as α-helices and β-sheets (Gregoret & Cohen, 1991; Hao *et al.*, 1992; Socci *et al.*, 1994; Hunt *et al.* 1994; Kolinski & Skolnick, 1992). Our analysis shows, however, that despite wide variation in the absolute number of residues in secondary structure, the stabilization free energy of secondary structures provided by compactness is essentially *independent* of the criterion used. Calculation of the stabilization free energy realized from compactness is estimated to be of the order of 2 kT per secondary structural unit.

Several energy minimization strategies using the AMBER potential are used to determine how far the maximally compact poly-alanine chains are from realistic energy minima. We find that small energetic perturbations to nearby local minima increased the number of dihedral angles in regions of ϕ/ψ space that are commonly observed in real proteins and nudged helices to become α-helices.

Compactness appears to be a structurally non-specific entropic force that lowers the overall conformational free energy for a large class of helix-like and strand-like structures, among which are the α-helices and β-strands that are specific to peptide backbones. We believe that other polymers could also be driven by compactness to adopt helical and sheet structures. But the microscopic geometric details would be dictated by their preferred backbone conformations. Many crystal structures of synthetic polymers, indeed, adopt helical or planar zig-zag conformations (Tadokoro, 1979). The stabilization that results from compactness is due to the vast reduction in the number of conformations of the chain that are accessible in compact states, due to excluded volume. This reduction is a process in which a large fraction of the remaining conformations contain helix-like and strand-like elements that are capable of filling space more densely than non-repeating conformations can.

We thank Sarina Bromberg and Nathan Hunt for helpful discussions, Hunt *et al.* and Socci *et al.* for making their manuscripts available prior to publication, and the NIH for financial support (grant numbers GM-34993 for the UCSF group and GM-38221 for T.H.).

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Seivers, R., eds), pp. 107–132, Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.

Anderson, A. G. & Hermans, J. (1988). Microfolding: conformational probability map for the alanine dipeptide in water from molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **3**, 262–265.

Chan, H. S. & Dill, K. A. (1989). Compact polymers. *Macromolecules*, **22**, 4559–4573.

Chan, H. S. & Dill, K. A. (1990*a*). Origins of structure in globular proteins. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 6388–6392.

Chan, H. S. & Dill, K. A. (1990*b*). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118–3135.

Chan, H. S. & Dill, K. A. (1991). Sequence space soup of proteins and copolymers. *J. Chem. Phys.* **95**, 3775–3787.

Chothia, C. (1975). Structural invariants in protein folding. *Nature (London)*, **254**, 304–308.

Dev, S. B. (1987). Quantitative prediction of protein secondary structure—where is the lacuna? *J. Biol. Phys.* **15**, 57–61.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.

Flory, P. J. (1949). The configuration of real polymer chains. *J. Chem. Phys.* **17**, 303–310.

de Gennes, P. G. (1979). In *Scaling Concepts in Polymer Physics*, pp. 54–61, Cornell University Press, Ithaca.

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

Hao, M. H., Rackovsky, S., Liwo, A., Pincus, M. R. & Scheraga, H. A. (1992). Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 6614–6618.

Havel, T. F. (1990). The sampling properties of some distance geometry algorithms applied to unconstrained computed conformations. *Biopolymers*, **29**, 1565–1585.

Havel, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Progr. Biophys. Mol. Biol.* **56**, 43–78.

Head-Gordon, T., Head-Gordon, M., Frisch, M. J., Brooks, C. L., III & Pople, J. A. (1991). Theoretical study of blocked glycine and alanine peptide analogues. *J. Amer. Chem. Soc.* **113**, 5989–5997.

Honig, B., Sharp, K. & Yang, A. (1993). Macroscopic models of aqueous solutions: biological and chemical applications. *J. Phys. Chem.* **97**, 1101–1109.

Hunt, N. G., Gregoret, L. M. & Cohen, F. E. (1994). The origins of protein secondary structure: effects of packing density and hydrogen bonding studied by a fast conformational search. *J. Mol. Biol.* **241**, 312–316.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karplus, M. & Shakhnovich, E. (1992). Protein folding: theoretical studies of thermodynamics and dynamics. In *Protein Folding* (Creighton, T. E., ed.), pp. 127–195, Freeman, New York.

Kolinski, A. & Skolnick, J. (1992). Discretized model of proteins I. Monte Carlo study of cooperativity in homopolymers. *J. Chem. Phys.* **97**, 9412–9426.

Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **114**, 181–239.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Pettitt, B. M. & Karplus, M. (1988). Conformational free energy of hydration for the alanine dipeptide: thermodynamic analysis. *J. Chem. Phys.* **92**, 3994–3997.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–483.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1–14.

Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Struct. Funct. Genet.* **3**, 71–84.

Socci, N. D., Bialek, W. S. & Onuchic, J. N. (1994). Properties and origins of protein secondary structure. *Phys. Rev. E.* **49**, 3440–3443.

Sosnick, T. R., Mayne, L., Hiller, R. & Englander, S. W. (1994). The barriers in protein folding. *Nature: Struct. Biol.* **1**, 149–156.

Tadokoro, H. (1979). *Structure of Crystalline Polymers*, Wiley, New York.

Tobias, D. J. & Brooks, C. L., III (1992). Theoretical study of blocked glycine and alanine peptide analogues. *J. Phys. Chem.* **96**, 3864–3870.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Jr & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Amer. Chem. Soc.* **106**, 765–784.

Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986). An all atom forcefield for simulations of proteins and nucleic acids. *J. Comput. Chem.* **7**, 230–252.

Yee, D. P. & Dill, K. A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci.* **2**, 884–899.

Edited by B. Honig

(Received 22 January 1994; accepted 24 May 1994)