Method for Predicting the State of Association of Discretized Protein Models. Application to Leucine Zippers†

Michal Vieth,‡ Andrzej Kolinski,§,|| and Jeffrey Skolnick*,§

Departments of Chemistry and Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037, and Department of Chemistry, University of Warsaw, Pasteura 1, 02-276 Warsaw, Poland

Received August 31, 1995; Revised Manuscript Received November 13, 1995⊗

ABSTRACT: A method that employs a transfer matrix treatment combined with Monte Carlo sampling has been used to calculate the configurational free energies of folded and unfolded states of lattice models of proteins. The method is successfully applied to study the monomer−dimer equilibria in various coiled coils. For the short coiled coils, GCN4 leucine zipper, and its fragments, Fos and Jun, very good agreement is found with experiment. Experimentally, some subdomains of the GCN4 leucine zipper form stable dimeric structures, suggesting the regions of differential stability in the parent structure. Our calculations suggest that the stabilities of the subdomains are in general different from the values expected simply from the stability of the corresponding fragment in the wild type molecule. Furthermore, parts of the fragments structurally rearrange in some regions with respect to their corresponding wild type positions. Our results suggest for an Asn in the dimerization interface at least a pair of hydrophobic interacting helical turns at each side is required to stabilize the stable coiled coil. Finally, the specificity of heterodimer formation in the Fos-Jun system comes from the relative instability of Fos homodimers, resulting from unfavorable intra- and interhelical interactions in the interfacial coiled coil region.

Since the coiled coil motif was first proposed by Crick (to explain the keratin diffraction pattern) (Crick, 1953, 1952), coiled coils have been the subject of increasing attention (Cohen & Parry, 1986; Cohen & Parry, 1990; Harrison, 1991; O'Shea *et al.*, 1991). Coiled coils consist of two or more helices wound around one another and can be found in the muscle protein tropomyosin (Johnson, 1975; Phillips, 1986), in blood clots as fibrin, and in hair as keratin (Fraser & MacRae, 1971; Cohen & Parry, 1986; Cohen & Parry, 1990). Furthermore, coiled coils play an important role in transcriptional activators (Landshultz *et al.*, 1988; Harrison, 1991) regulating cell growth, differentiation, and oncogenesis and therefore are important medically (Perutz, 1992). In these proteins, coiled coils, named leucine zippers because they possess a Leu in every seventh position, have been proposed to form dimerization domains (Landshultz *et al.*, 1988; Harrison, 1991). The biological importance and relative simplicity of leucine zippers have made them the subject of extensive experimental (Hodges *et al.*, 1981; O'Shea *et al.*, 1989, 1991, 1993; Smeal *et al.*, 1989; Harbury *et al.*, 1993; Lovejoy *et al.*, 1993; Lumb *et al.*, 1994) and theoretical studies. They have been used as models for studying various interactions responsible for driving protein folding (Harbury *et al.*, 1993; O'Shea *et al.*, 1993), for testing hypotheses about the specificity and stability of protein structures (Harbury *et al.*, 1993; O'Shea *et al.*, 1993), and for understanding factors influencing oligomeric assembly and protein design principles (Harbury *et al.*, 1993; Lovejoy *et al.*, 1993; O'Shea *et al.*, 1993). Because of the abundance of interesting experimental data and their small size and relative simplicity, leucine zippers are ideal model systems for various theoretical approaches to the protein folding problem (Krystek *et al.*, 1991; Nilges & Brunger, 1993; Zhang & Hermans, 1993; DeLano & Brunger, 1994; Vieth *et al.*, 1994a). Taking advantage once again of the simplicity of leucine zippers, the present work focuses on developing a theoretical approach to understanding the factors responsible for the stability of dimeric coiled coils relative to the monomeric structures.

The sequences of coiled coils have a characteristic heptad repeat $(abcdefg)_n$ (McLachlan & Stewart, 1975). Residues at positions a and d are in general occupied by hydrophobic residues and form an interface between two or more helices (in leucine zippers, position d is occupied by Leu residues). The e and g positions are occupied mostly by charged residues, and the methyl groups of those residues form the edges of hydrophobic core. The b, c, and f positions are mostly hydrophilic. In addition to having Leu in most d positions, leucine zipper fragments of DNA binding proteins are short (23−45 residues long) (Landshultz *et al.*, 1988), and some of them, when excised from their parent proteins, have a tendency to dimerize (O'Shea *et al.*, 1989, 1991). A number of transcription factors form heterodimers, e.g., c-Fos and c-Jun (Hai *et al.*, 1989; Smeal *et al.*, 1989), with leucine zipper interfaces. The isolated leucine zipper domains of c-Fos and c-Jun also have a tendency for heterodimerization (O'Shea *et al.*, 1989), similar to the intact proteins.

The experimental observation that attracted our attention was the phenomenon of subdomain folding in the GCN4 leucine zipper system (Lumb *et al.*, 1994). The various subdomains (fragments) of the wild type GCN4 leucine zipper were synthesized, and their thermal stability was assessed. The fragment corresponding to residues 8−30 of

† This work was supported in part by NIH Grants GM-38794 and 1R03-TW00418.

* Corresponding author. Phone: (619) 554-4821; Fax: (619) 554-6717.

‡ Department of Chemistry, The Scripps Research Institute.

§ Department of Molecular Biology, The Scripps Research Institute.

|| University of Warsaw.

⊗ Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0006-2960/96/0435-955$12.00/0

the wild type GCN4 leucine zipper was found to have substantial stability. In contrast, the closely related fragment corresponding to residues 11−33 in the wild type GCN4 was predominantly unfolded. In order to investigate this problem, we need to develop a methodology to extract from simulations the equilibrium constants for formation of coiled coils from unfolded monomeric chains. Previously, the prediction of the folding pathways and structure of the GCN4 leucine zipper has been reported by our group (Vieth *et al.*, 1994a). Subsequently, the calculation of the oligomerization equilibria of GCN4 leucine zipper and some of its mutants was then described (Vieth *et al.*, 1995). However, unfolded monomeric chains cannot be treated by this method. Therefore, in this paper, a new more general approach is proposed to calculate free energies of the unfolded as well as folded chains.

The outline of the paper is as follows. In the Method section, the general expression for the equilibrium constant of monomer−dimer equilibria is presented. The treatment of the equilibria between monomers and dimers presented in this paper is in principle general and not limited to the lattice models of proteins. However, calculation of different contributions to the free energy is presented in the context of a lattice model of proteins (Kolinski & Skolnick, 1994; Vieth *et al.*, 1995a). Thus, the lattice model of proteins is briefly introduced and the calculation of different contributions to the dimerization free energy is described. In particular, a transfer matrix treatment is presented to evaluate the internal contributions to the free energy of dimerization. In the appendices, we present the detailed description of the treatment as well as our efforts to validate the methodology. The Results section focuses on the calculation of the dimerization free energy for the GCN4 leucine zipper and its fragments, the monomer−dimer equilibrium in Fos, and the heterodimer−homodimer equilibria in the Fos-Jun system. In the Discussion section, the implications of our results are presented, together with elaboration of the necessary condition for stabilization of single hydrophilic residues in the helical interface of coiled coils.

METHOD

In what follows, we present an overview of the calculation of the monomer−dimer equilibria for a pair of chains. We first present the general statistical mechanical formalism. Then, the lattice model of proteins is introduced and the calculation of different components of free energy of dimerization are described.

Formalism for Monomer−Dimer Equilibria

Consider the equilibrium constant for dimerization (McQuarrie, 1976):

$$2\mathrm{M} \leftrightarrow 1\mathrm{D} \tag{1}$$

$$K = [\mathrm{D}]/[\mathrm{M}]^2 \tag{2}$$

with [M] and [D] being the equilibrium concentrations of monomers and dimers, respectively. The total concentration of individual chains C_0 (assuming that the only species in the system are monomers and dimers) is expressed as:

$$C_0 = 2[\mathrm{D}] + [\mathrm{M}] \tag{3}$$

whereas the fraction of individual chains in monomers and dimers is given by:

$$x_\mathrm{M} = \frac{[\mathrm{M}]}{C_0} \qquad x_\mathrm{D} = \frac{2[\mathrm{D}]}{C_0} \tag{4}$$

Obviously, $x_\mathrm{M} + x_\mathrm{D} = 1$. Substituting x_M and x_D into eq 2:

$$K = \frac{x_\mathrm{D}}{2C_0 {x_\mathrm{M}}^2} \tag{5a}$$

and:

$$x_\mathrm{M} = \frac{-1 + \sqrt{1 + 8KC_0}}{4KC_0} \tag{5b}$$

Thus, the calculation of x_M and x_D require values for both K and C_0. In order to relate the equilibrium constant K to the microscopic variables, it is useful to rewrite eq 5 in the form of mole fraction equilibrium constant K^x (McQuarrie, 1976):

$$K^x = C_0 K = \frac{x_\mathrm{D}}{2{x_\mathrm{M}}^2} \tag{6a}$$

Substituting the total number of chains for the mole fractions:

$$K^x = \frac{2N_\mathrm{D}/N}{2(N_\mathrm{M}/N)^2} = N\frac{N_\mathrm{D}}{{N_\mathrm{M}}^2} \tag{6b}$$

where N is the total number of individual chains in a box of volume V, and N_M and N_D are numbers of monomers and dimers, respectively. $N_\mathrm{D}/{N_\mathrm{M}}^2$ is simply the ratio of the total partition function for dimers Z_D divided by the square of the total monomer partition function Z_M. Thus, K^x can be written in the following form (McQuarrie, 1976; Skolnick, 1980):

$$K^x = N\frac{Z_\mathrm{D}}{{Z_\mathrm{M}}^2} = N\frac{VV_0Z_{\mathrm{conf,D}}}{\sigma_\mathrm{D}(VZ_{\mathrm{conf,M}})^2} \tag{7a}$$

$Z_{\mathrm{conf,M(D)}}$ are the configurational partition functions for the monomers, M (dimers, D), that include the rotational and internal contribution. V_0 is the volume accessible to a first atom in the second chain in the dimer, given that the first atom in the first chain of the dimers is fixed, σ_D is the symmetry number (=2! for homodimers and 1 for heterodimers). Since we choose an internal coordinate system (fixed at the first bead of chain one) to calculate the partition function, the factor V comes from the integration over the degrees of freedom of the first bead in the first chain. As noted in a series of papers, the contributions from the momenta degrees of freedom cancel in the numerator and denominator (Herschbach, 1959; Mayer & Mayer, 1963; Holtzer, 1995) (for this internal coordinate system), leaving the ratio of configurational partition functions to determine the equilibrium constant. Rearranging eq 7a, we get

$$K^x = \frac{V_0}{V_\mathrm{M}}\frac{Z_{\mathrm{conf,D}}}{\sigma_\mathrm{D}{Z_{\mathrm{conf,M}}}^2} \tag{7b}$$

Let us note that $V_\mathrm{M} = N/V$ is the average volume accessible per chain and is equal to $1/C_0$.

Equations 6 and 7b provide the basic equations for evaluation of the fraction of monomers and dimers in the system at a given concentration C_0. In order to dissect the

total contribution to the free energy change for the reaction in equation 1, let us define the standard molar free energy change for the monomer−dimer equilibrium (McQuarrie, 1976):

$$\Delta A^{\theta}_{M\rightarrow D} = -kT \ln K^{x} \quad (8a)$$

Combining 8a with eq 7b:

$$\Delta A^{\theta}_{M\rightarrow D} = -kT \ln (V_0/V_M) - kT \ln Z_{conf,D} + 2kT \ln Z_{conf,M} + kT \ln \sigma_D \quad (8b)$$

The total free energy of dimerization can be written as a sum of a translational and a configurational part:

$$\Delta A^{\theta}_{M\rightarrow D} = -T\Delta S_{tr}(V_M) + \Delta A^{\theta}_{conf} \quad (8c)$$

with $\Delta S_{tr} = k \ln (V_0/V_M)$ and $\Delta A_{conf} = \Delta A_{int} - T\Delta S_{rot}$. The total free energy change from eq 8a−c can be viewed as three subprocesses corresponding to three different energy contributions (translational, rotational, and internal). Figure 1a−d shows a schematic thought experiment that can be done to understand the meaning of different contributing terms to the free energy of dimerization. The first subprocess (Figure 1b) can be viewed as the loss of translational degrees of freedom upon bringing the two first chains together. The methodology for calculating the translational entropy is shown in Figure 2. Next, an end is oriented, and rotational entropy is lost (Figure 1c). Finally, the native interactions are formed (Figure 1d). In the following, a method to calculate each contribution to the free energy for the lattice model of proteins is described.

Lattice Model of Proteins

Since the details of the lattice model have been presented elsewhere (Kolinski & Skolnick, 1994; Vieth *et al.*, 1995a), we present here only a brief description of the methodology. Additional, salient details are provided in Appendix A. The protein is modeled by an α-carbon representation of the backbone and by single sphere, multiple rotamer side chains. The α-carbons are connected by a set of vectors of the type 1.22*{(2, 2, 0), (2, 2, 1), (3, 0, 0), (3, 1, 0), (3, 1, 1)} (Kolinski & Skolnick, 1994; Vieth *et al.*, 1995a). The entire statistical interaction scheme, derived from high resolution structures (see the sample derivation in the recent paper by Godzik *et al.* (1995)) from the Protein Data Bank (Bernstein *et al.*, 1977), is subdivided into short and long range interactions. Among the short range interactions, there are local conformational preferences of neighboring residues along the sequence together with a rotamer energy. By the term long range interactions, we mean all interactions between residues that are at least 4 residues apart in sequence. All long range interactions are sequence dependent. They consist of a side chain pair potential, a cooperative pair potential (which facilitates interactions between secondary structural elements, i.e., helices and β-sheets), and a contact based on body term (similar to a solvation energy). The model hydrogen bonds (derived in the spirit of Levitt and Greer (1977)) can be local (helical) or long range (in β-sheets) and operate only between α-carbons. The conformations are sampled using a Monte Carlo procedure (Metropolis, 1953).

Configurational Partition Function for the Denatured State. In the unfolded state, both chains are noninteracting

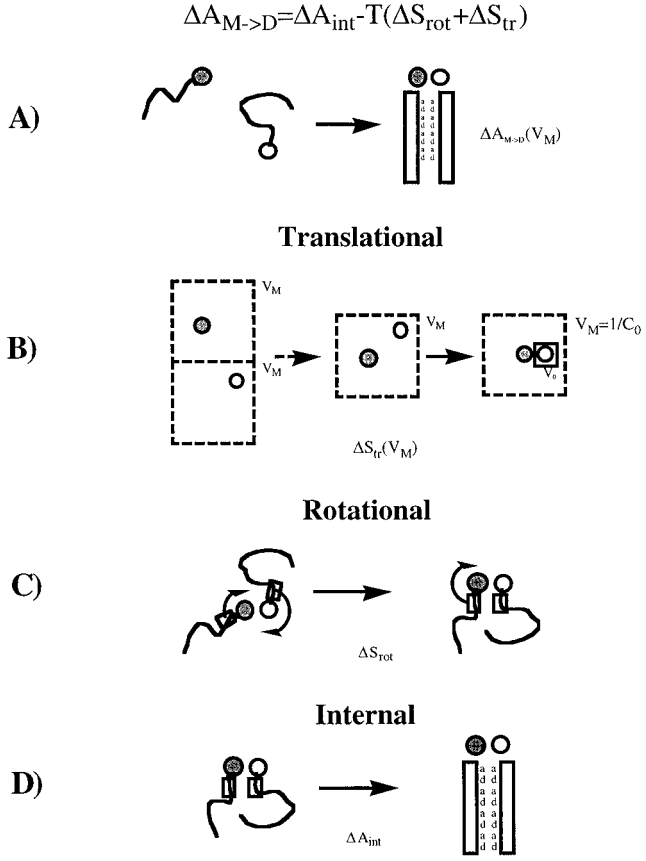

FIGURE 1: A thought experiment that dissects the different contributions to the free energy of dimerization. The shaded circle represents the first bead of the reference chain (the first chain). The open circle represents the first bead of the second chain. The first chain has freedom to translate and rotate in both the unfolded and folded state. (A) The entire dimerization process under consideration. On the left-hand side are the reactants—two independent chains—and on the right-hand side are the products—a two chain, parallel coiled coil with specific interactions. (B) Reduction in translational entropy associated with bringing the first bead of the second chain (open circle) close to the first bead of the first chain (shaded circle). On the left-hand side, both beads have an average accessible volume V_M (V_M is the average volume per molecule, $1/C_0$). After they are brought together, the first bead has a volume V_M accessible to it, whereas the second bead has only a small vibrational volume V_0 in the dimer. (C) The change in rotational part of configurational free energy. The left-hand side shows free rotation of both chains, whereas on the right-hand side the two chains can rotate only as a unit. The relative rotation of the second chain with respect to the first one is restricted. (D) The change in the internal free energy. On the left-hand side, the two chains have a fixed relative orientation (determined by the first two vectors in both chains) but do not interact and can assume any conformation. On the right-hand side, a two chain, coiled coil with specific interchain interactions is formed.

and, thus, can be treated independently. For short chains (containing less than 6 bonds), it is possible to do an exact enumeration of all possible states to obtain the total internal partition function for the unfolded state. Unfortunately, for the longer chains, the exact enumeration is computationally intractable. However, if one assumes that the total energy of a chain in a particular conformation can be written as a sum of energies of small overlapping segments (i.e., 4 bond segments), then the partition function can be calculated using a transfer matrix treatment (Zimm & Bragg, 1959; Lifson & Roig, 1961; Flory, 1969; Poland & Scheraga, 1970) described in detail in Appendix B. This type of treatment was extensively used in helix−coil transition theory (Zimm

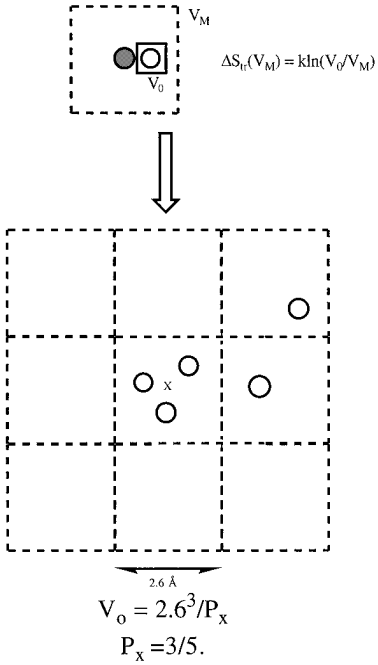

FIGURE 2: Schematic representation of calculation of loss of translational entropy for the first (white) bead of the second chain. In the folded state, the white bead has an accessible vibrational volume V_0. The bottom picture shows a method of calculating this volume. The coordinates of the first bead of the second chain are first expressed in the coordinate system centered at the first bead of the first chain. Note that the entire space is discretized and divided into small cubic boxes of a side of 2.6 Å. Then, the simulation of folded state is run, and the number of occurrences of the white bead is counted in each box every 100 MC cycles. The maximally occupied box is selected, and the frequency of the white bead being in this box is calculated. For example, in this case we have 5 snapshots, and the number of times the maximally occupied box is visited is 3. The vibrational volume V_0 is computed as the ratio of the volume of the elementary box (2.6 Å^3) to the probability of a white bead being in the maximally occupied box ($P_x = {}^3/_5$).

& Bragg, 1959; Lifson & Roig, 1961; Poland & Scheraga, 1970) as well as in polymer physics (Flory, 1969).

The approach we use in this paper is a generalization of the treatment proposed by Skolnick and Kolinski (1992). The entire polypeptide chain is divided into N-4 four bond (five residue) overlapping segments. Each segment is treated as being in one of several discrete rotational states (Flory, 1969). Only interactions within each segment are considered. For each of the overlapping four bond segments, we calculate the statistical weights of all possible conformations and use these weights to build N-4 transfer matrices. Multiplication of these matrices and summation of the resulting elements give the partition function for the system.

This treatment neglects the long distance excluded volume effects, as well as the statistical weight of interacting clusters of residues, which would be created if the chain folded onto itself (Flory, 1969). Local elements of secondary structure (turns and helical states) are accounted for in this treatment. For short chains, long range, repulsive excluded volume effects and the attractive contributions from hydrophobic clusters on average cancel and yield free energies which are close to those obtained by the exhaustive enumeration. The exact enumeration was compared to the transfer matrix treatment for six (-Gly-Ala$_4$-Gly-) and seven residue (-Gly-Ala$_5$-Gly-) chains. For the six residue chain, both treatments give identical results with free energies $-14.2\ kT$. For the seven residue chain (-Gly-Ala$_5$-Gly-), the exact enumeration gives a slightly lower free energy ($-17.1\ kT$) than the transfer matrix treatment ($-16.3\ kT$) does, but both values are close enough to be considered similar. For longer chains, it is unclear whether the cancellation in the free energy calculation persists, but we use this method as a first approximation to the free energy of the unfolded state. The contribution of the long range interactions to the unfolded state free energies for longer chains (up to 11 residues) is being evaluated; however, any speculation on this issue is beyond the scope of this paper.

Configurational Partition Function for Dimeric Molecules. The folded state is considered to be a subset of conformations that have a specific overall topology, hydrogen bond pattern, and specific side chain contact map (Ptitsyn, 1987). The entire configurational partition function for the folded state can be written as a product of the rotational and internal parts (McQuarrie, 1976):

$$Z_{\text{conf,D}} = Z_{\text{rot}} Z_{\text{int,D}} = N_{\text{D1}} N_{\text{D2}} \sum_{i=1}^{n_{\text{levels}}} n(E_i) \mathrm{e}^{-\beta E_i} \qquad (9)$$

N_{D1} is the number of distinct conformational states for the first two vectors in the first chain (assumed to be equal 4704, as determined by statistics), whereas N_{D2} is the number of distinct states for the first two residues in the second chain. $n(E_i)$ is the degeneracy of energy level E_i, n_{levels} is the number of energy levels, and $\beta = 1/kT$. The probability of being in any energy level (for convenience, we choose the average energy level E_0) is given by (McQuarrie, 1976):

$$P(E_0) = \frac{n(E_0) \exp(-\beta E_0)}{Z_{\text{int}}} \qquad (10)$$

$P(E_0)$ is the probability of the system being in E_0. If we could get an estimate for the degeneracy of one energy level (say $n(E_0)$ for convenience), then we would be able to obtain the configurational partition function for the system:

$$Z_{\text{conf,D}} = N_{\text{D1}} N_{\text{D2}} \frac{n(E_0)}{P(E_0)} \exp(-E_0/kT) \qquad (11)$$

We can now write the expression for the configurational free energy of the folded state:

$$A_{\text{conf,D}} = -kT \ln Z_{\text{conf,D}} = -kT \ln N_{\text{D1}} N_{\text{D2}} + E_0 - kT \ln n(E_0) + kT \ln P(E_0) \qquad (12)$$

where $k \ln(N_{\text{D1}} N_{\text{D2}})$ is the rotational entropy term, and $k \ln(n(E_0)) - k \ln(P(E_0))$ is the internal entropy of the system arising from the degeneracy of the average energy level ($k \ln(n(E_0))$ and the fluctuation terms ($-k \ln(P(E_0))$).

The average energy of the system together with the energy probability distribution (Figure 8) can be obtained from Monte Carlo simulations of the folded state (see Appendix C). First, we generate a starting structure for each sequence in a parallel coiled coil arrangement. Then, production runs are carried out to calculate the average energy of the system together with the energy probability distribution. The degeneracy of the energy level E_0 $n(E_0)$ is calculated from the ensemble of structures within 4 kT of the average energy using a similar transfer matrix treatment as for the unfolded state.

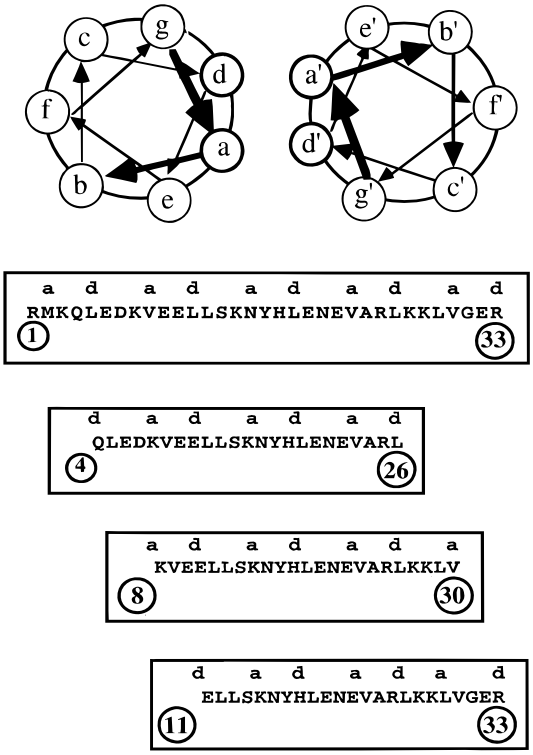

FIGURE 3: Schematic depiction of the fragments. The top layer shows a diagram of a helical wheel with the heptad residue repeat indicated. The rectangles at the bottom show the sequences of the GCN4 wild type and all peptide fragments investigated in this study together with assignment to the hydrophobic heptad positions a and d. Note that each peptide fragment starts with a residue adjacent to the hydrophobic core and ends with a residue that would be in the core if the helices were longer.

RESULTS

Using the procedure described in the Method section, the free energies of the unfolded states and the free energies of folded sequences in the structure of double stranded, parallel coiled coils have been calculated. In all cases studied here, we eliminated the possibility of formation of antiparallel species by energetic considerations (see fitness function in Figure 7). A number of sequences have been investigated. First, there is the GCN4 leucine zipper wild type, shown experimentally to be a dimeric coiled coil (O'Shea *et al.*, 1991). Next, there are various 23 residue fragments (subdomains) of the GCN4 leucine zipper studied by Lumb *et al.* (1994) indicated schematically in Figure 3. Each subdomain starts from the residue adjacent to the hydrophobic core (at the g and c positions) and ends at the position that was in the hydrophobic core in the wild type. We expect that, at both ends, the residues that initiate or terminate the hydrophobic core will be destabilized with respect to the identical residues in the wild type. The reason for that is loss of interactions with the preceding or subsequent residues (present in the wild type but absent in the fragment) and partial exposure of those residues to solvent.

The fragment corresponding to residues 8−30 of the wild type GCN4 was found to be stable as a coiled coil dimer, whereas the closely related fragment corresponding to residues 11−33 of GCN4 wild type appeared to be predominantly unfolded (Lumb *et al.*, 1994). Lumb *et al.* (1994) concluded that specific packing interactions can be energetically more important than local secondary structure propensities. Lumb *et al.* also suggested that the protein folding can be understood in terms of formation of cooperatively folded domains. In addition to these two peptides, we also test another 23 residue fragment corresponding to the residues 4−26 of the wild type GCN4, which has not yet been studied experimentally. However, based on the energetic profile of the wild type (Vieth *et al.*, 1995a), we would expect the 4−26 fragment to form the most stable 23 residue dimeric coiled coil.

Monomer−Dimer Equilibria in the GCN4 Leucine Zipper. Over the experimental concentration range, the predicted fraction of chains in the dimeric structure is shown in Table 1 for the GCN4 leucine zipper. Because of the limited accuracy of our free energy estimates, we restrict ourselves to the assignment of the dominant species. Our results are in agreement with experiment and indicate that the wild type of GCN4 leucine zipper should populate dimeric chains. Table 1 also shows the calculated conformational (rotational parts included) free energies of the GCN4 leucine zipper in the unfolded state as well as in the coiled coil structure. The free energy changes upon folding are also presented. The translational entropy loss upon dimer formation is shown in the second last column for three different concentrations (2 μM, 43 μM, 1 mM) (Lumb *et al.*, 1994). In addition, Table 1 presents the average RMS fluctuations from the average structure in the coiled coil state together with the volume occupied by the first bead of the second chain.

In order to evaluate the energetic and entropic contributions to the free energy of the unfolded state, unrestrained Monte Carlo simulations of GCN4 leucine zipper wild type monomer (in $T = 1.85$ with no long range interactions) have been performed. The difference between the free energy computed by the transfer matrix approach (see Appendix B) and the average energy computed from the Monte Carlo simulation was considered to be the entropy of the unfolded state. Table 1 presents the dissection of the free energy of folded and unfolded state of the wild type GCN4 leucine zipper. As would be expected, the dominant contribution to the unfolded state configurational free energy comes from the entropy. The average energetic contribution is small, and by our assumption, there is no energetic contribution from the long range interactions. Upon folding, this situation changes dramatically. The dominant contribution to the free energy of the folded state comes from the energy (−96.6 kT per monomer), and the entropy drops down by roughly 50% in comparison to the unfolded state. The major energetic contribution to the folded state comes from hydrogen bonds, local side chain orientational preferences, pairwise interactions, and cooperative side chain packing interactions.

From the values of the internal entropy presented in Table 1, we can calculate the average number of states per residue for the unfolded and the folded state. For the unfolded state, the average number of states per residue is close to 45. For the folded state, this number decreases to 7. Thus, our model provides almost a 7-fold reduction in the number of states upon folding. This reduction is slightly lower than the 8-fold reduction estimated by Privalov (1992). This relatively minor difference in entropy change can be due to the fact that coiled coils have larger exposed surface area than globular proteins as well as to approximations of our approach. It is noteworthy that if the sequence of the GCN4 leucine zippers were in the tetrameric state, then the calculated reduction in the average number of states per residue upon folding would be roughly 7.7; this is very close to the number estimated by Privalov. This may not be

Table 1: Thermodynamical and Structural Parameters for the GCN4 Leucine Zipper

dominant species[a]		free	monomer free	dimer free	translational entropy loss[h]			RMS
43 μM	1 mM	energies	energy dissection	energy dissection	2 μM	43 μM	1 mM	in Å[i]
2	2	−149.1[b]	−14.8 (0)[e]	−96.6 (41)[e]				1.7
		−336.3[c]	125.8[f]	64.3[f]	16.3	13.3	10.1	(1.3)
		−38.1[d]	45[g]	7[g]				

[a] The predicted dominant species for concentrations 43 μM and 1 mM ("2" indicates that dimers only are present) are shown. Experimentally, only dimers are present at these concentrations. [b] The configurational free energy of the unfolded state monomers in *kT*. [c] The configurational free energy of the folded state (coiled coil dimer) in *kT*. [d] The configurational free energy change upon folding (the free energy of dimer minus twice the monomer free energy) in *kT* units. [e] The average energy per chain in *kT* (percentage (%) of long range interactions is shown in parentheses). [f] The configurational entropy per chain multiplied by the reduced temperature (in *kT*). [g] The average number of states per residue. [h] The translational entropy loss upon formation of the dimer multiplied by the reduced temperature for different concentrations (2 μM, 43 μM, 1 mM) in *kT*. The average volume occupied by the first bead of the second chain V_0 is 67.6 Å^2. [i] The average C_α RMS deviation of a single structure from the average structure (standard deviation is shown in parentheses).

Table 2: Comparison with Experiment of the Predicted Dominant Species for GCN4 Leucine Zipper Fragments

	concn dependence of monomer/dimer ratio		dominant species[a]	
protein	43 μM	1 mM	theory	expt
GCN4 8−30	0:100	0:100	2	2
GCN4 11−33	100:0	99:1	1	1
GCN4 4−26	9:91	2:98	2	NA

[a] "2" ("1") indicates the presence of dimers (monomers) only, "NA" indicates that experimental data are not available.

Table 3: Free Energy of the Unfolded State and the Folded State (Coiled Coil Parallel Dimer) for Different Fragments of the GCN4 Leucine Zipper

protein	free energy of unfolded chain	free energy of folded chain	$\Delta A^{\theta a}_{conf}$	translational entropy loss[b]		
				2 μM	43 μM	1 mM
8−30	−101.1	−228.6	−26.4	17.0	13.9	10.8
11−33	−103.1	−209.8	−3.6	14.0	10.9	7.8
4−26	−101.9	−222.0	−18.2	16.5	13.4	10.3

[a] ΔA^{θ}_{conf} is the configurational free energy change upon folding including loss of rotational entropy $\sim kT \ln(N_{D1}/N_{D2})$. [b] Translational entropy loss, $-k \ln(V/V_0)$ multiplied by temperature.

Table 4: Structural Parameters for the Dimers of the GCN4 Leucine Zipper Fragments Computed from Monte Carlo Simulations

	av RMS values for structures		
protein	for all structures[a]	for structures within 4*kT* from the av energy[b]	vol occupied by the first bead V_0[c]
8−30	1.0 (0.3)	1.0 (0.3)	35.08
11−33	3.4 (1.1)	3.2 (0.9)	680.2
4−26	1.1 (0.5)	1.1 (0.5)	57.1

[a] Average RMS (standard) deviations from the average structures for all structures occurring in the simulation. [b] Average RMS deviations from the average structures computed from those structures within 4*kT* of the average energy. [c] Volume occupied by the first bead V_0 in Å^3 of the second chain (provided that the first bead of the first chain is pinned) and averaged over all simulations.

surprising, since coiled coil tetramers closely resemble the four helix bundle topology observed in globular proteins. What is surprising, however, is that the number of states per residue in the folded state is so large. Most of the variation that we observe comes from the mobility of the side chains in the exposed parts of the molecule as well as in the exaggerated backbone mobility on the lattice. The idea of many conformational substates (distinguished by small changes in the protein structure) in the folded state was widely elaborated on by Frauenfelder *et al.* (1988). Nevertheless, we feel that our model overestimates the number of these states; it also overestimates the range of conformational fluctuations in the folded state. As a reminder, let us note that the inherent resolution of our lattice imposed by its geometry and wide interaction basins is estimated to be around 2.6 Å (Kolinski & Skolnick, 1994). Also the average RMS of C_α atoms with respect to the average structure for the folded state is around 2 Å, with fluctuations ranging from 0.8 to 3.5 Å (see Tables 1, 4, and 9). Thus, the lattice model has a folded state defined as a tube with a 1.3 Å radius; this is probably why such a large number of states per residue in the native conformation is seen.

Equilibria for GCN4 Leucine Zipper Subdomains. Tables 2−6 summarize the results for the fragments of the GCN4 leucine zipper. The predicted monomer/dimer ratio for different sequences is shown in Table 2. For all of the experimentally tested peptide fragments, the predicted dominant species are in agreement with experiment; i.e., fragment 8−30 is predicted to form a stable dimer, and 11−33 is predicted to be predominantly unfolded. For the fragment corresponding to residues 4−26 of the wild type, we predict the preferential formation of dimers.

Table 3 shows the free energies of the unfolded and folded states for the fragments. From these data one can infer which subdomain of the wild type GCN4 contributes mostly to the stability of the parent structure. Focusing on the configurational free energy contribution, the largest stability per residue is exhibited by the 8−30 fragment of the wild type ($\sim -1kT$ per residue). Another feature of the data from Table 3 is that the unfolded state free energies for the fragments are very close to one another and that the differences in stability reside in the folded state. Except for the 11−33 fragment (whose first 10 residues are disordered), the values of the translational entropy loss on dimerization are also similar for most tested sequences. Table 4 shows the average RMS fluctuations of a single structure from the average structure. The fluctuations of the 11−33 fragment are noticeably larger than in either the 4−26 or 8−30 fragments.

As noted above, the different stabilities of the fragments arise mostly from the differences in folded state free energies. Table 5 shows the dissection of the folded state free energies into the average energy and entropy. For the 4−26 and 8−30 fragments, which by our prediction appear as stable dimers, the entropic contributions are similar within the error of our calculations. For the 11−33 fragment that is predicted to be unfolded, the entropic contribution is larger. This relatively large entropic contribution of the folded state

Table 5: Dissection of Folded State Free Energy for the Fragments of GCN4 Leucine Zipper[a]

protein	energy	$-TS_{int}$
4−26	−129.0	−78.6
8−30	−137.5	−78.0
11−33	−92.3	−100.4

[a] All values shown are in *kT*. Rotational entropy TS_{rot} is 14.4*kT*, 13.1*kT*, and 17.1*kT*, for 4−26, 8−30, and 11−33 fragments, respectively.

Table 6: Comparison of the Estimated Configurational Free Energies of Dimer Formation (from the Wild Type per Residue Plot) and Calculated for Fragments

protein	estd ΔA^{θ}_{conf}	calcd ΔA^{θ}_{conf}
4−26	−28.1	−18.2
8−30	−22.7	−26.4
11−33	+2.7	−3.6

[a] Indicates the configurational free energy of dimer minus twice the free energy of monomers obtained from the wild type energy per residue values. [b] Indicates the configurational free energy of dimer minus twice the free energy of monomers calculated explicitly from the simulation and transfer matrix treatment. All values in *kT* units.

cannot compensate for a small energetic stabilization; thus, unfolded monomers are preferred. In agreement with the larger configurational entropy for the 11−33 fragment, the RMS fluctuations obtained from the Monte Carlo simulations, in Table 4, also show larger values. This is in agreement with the larger configurational entropy values for this fragment. Table 5 shows that the 8−30 fragment is (consistent with overall free energy values) energetically the most stable. The unstable fragment 11−33 shows a substantially lower energetic contribution than others. These data suggest that analysis of the differential stability of these fragments could be done based on the energetic considerations in the folded state.

Another interesting analysis bears on questions of the difference between the wild type and its constituent fragments. Could the different stabilities of fragments be rationalized from the plots of free energy of dimerization (the free energy of a dimer minus the free energy of two monomers) per residue using the wild type data? Are there any regions that in the fragments (apart from the ends likely to destabilize the molecules in comparison to the corresponding residues in the wild type) are substantially different in stability in comparison to the wild type? Table 6 shows the estimated values of the total free energy of dimerization assuming that all of the contributions for all of the residues are as in the wild type. The estimated values predict that fragment 11−33 be unstable, whereas 4−26 and 8−30 should exist as stable dimers, with the highest stability assigned to the 4−26. The only fragment for which the difference between the estimated and calculated values could be assigned to the end effects is 4−26. For this fragment, the calculated value is higher that the estimate. The difference could be assigned to the expected destabilization of the dimeric structure by partially unburied hydrophobic residues at both ends. The explicit calculations show that the 8−30 fragment is the most stable and is even more stable than would be estimated from the wild type data. This suggests possible additional stabilization of some residues with respect to the wild type which may arise from slightly different side chain packing. A similar but opposite effect is present in the least stable 11−33 fragment.

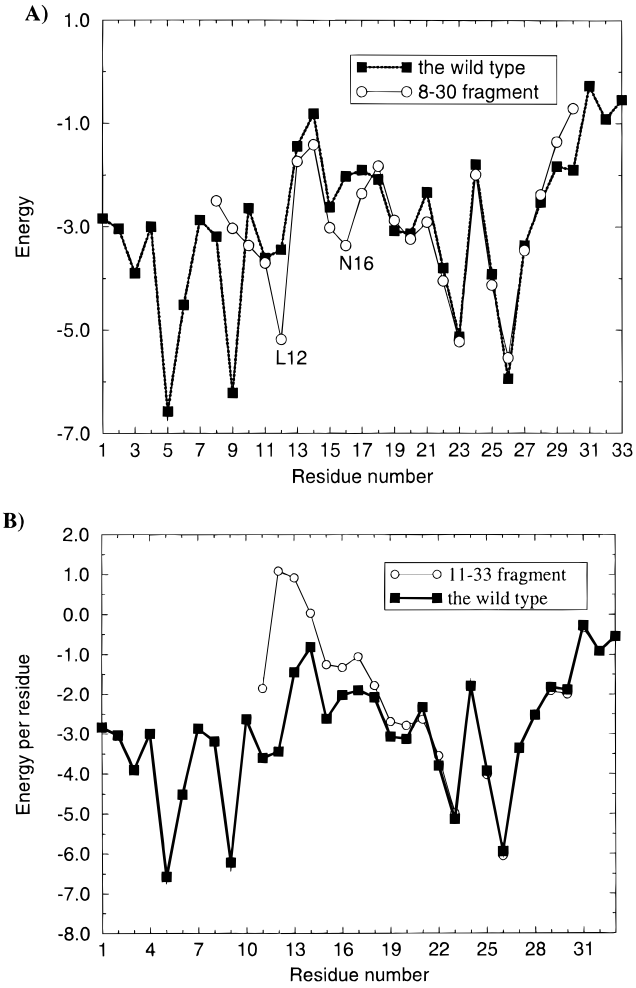

FIGURE 4: Energies per residue per chain of parallel coiled coil dimers for the 8−30 fragment (A) and 11−33 fragment (B) in comparison to energy per residue for the wild type (solid line). Fragment plots are shown as thin lines. The energy per residue is calculated as the sum of all of the energetic terms for a given residue, averaged over all of the simulations. Each energy term in which two (or more) residues participate is divided by two (or more), i.e., pair energy, hydrogen bond energy, and template energy. The sum of energies over all residues gives the total average energy of the system. (A) Plot showing the energy per residue in the 8−30 fragment (bold line with filled squares) in comparison to the wild type (thin line with open circles). Note the relative destabilization of the 8−30 fragment at the N- and C-terminus with respect to the wild type. Also residues at "old" 12 and 16 positions of a fragment experience additional stabilization relative to the wild type. (B) Plot of the energy per residue for the 11−33 fragment (bold line with filled squares) in comparison to the wild type (thin line with open circles). For 12 last residues, the two curves are identical. However, destabilization of the N-terminus in the fragment extends for the first 11 residues.

Figure 4A,B shows the energy per residue plots for two fragments 8−30 and 11−33 compared to the wild type. As seen in Figure 4A, the 8−30 fragment is, as expected, destabilized at both ends, probably because residues at the edges of hydrophobic core lose interactions with partners that used to be there in the wild type. However, substantial additional stabilization is exhibited by two residues in the helical interface, i.e., 5 Leu, which occupies the d position (12 Leu in the wild type) and 9 Asn at an a position (16 Asn in the wild type). By examining the plots of the various energy terms per residue, two types of interactions are found to be responsible for their additional stabilization (see Table 7). The largest stabilization comes from the pair interaction

Table 7: Dissection of the Extra Energetic Stabilization of the 8−30 Fragment by Residues 5 Leu and 9 Asn (d and a Positions)[a]

residue	estd pair energy	calcd pair energy	estd local E_β	calcd local E_β
5 Leu d	−4.6	−6.8	−1.3	−2.1
9 Asn a	0.34	−1.5	−1.1	−1.7

[a] All values in *kT* units per monomer.

energy, and a slightly smaller contribution originates from the local side chain orientational preferences. The pair energy difference comes from the slight side chain rearrangement and loss of interactions between 5 Leu and 9 Asn (this interaction is strongly repulsive in our statistical pair potential (Vieth *et al.*, 1995a)).

The RMS differences between the average structures for the fragments and the corresponding region in the wild type are 1.1, 1.5, and 2.1 Å for the 4−26, 8−30, and 11−33 fragments, respectively. These values are on the order of the RMS fluctuations for Monte Carlo simulations and do not preclude the possibility of side chain repacking.

In Figure 4B, the first 10 residues of fragment 11−33 show substantial energetic destabilization with respect to the wild type (Figure 4B) for the first 10 residues. Visual inspection of the Monte Carlo trajectories leads to the conclusion that first 10 residues are highly disordered. This is in agreement with quite large RMS fluctuations for the entire 11−33 fragment shown in Table 4. The residue that is responsible for this series of observations is Asn 16. A possible explanation is that the hydrophilic Asn in the wild type is forced to be in the hydrophobic core of the coiled coil structure by the stabilizing interactions, namely, cooperative hydrogen bonding and favorable hydrophobic interactions elsewhere in the molecule. The lack of an internal disordered region preceded and followed by interacting helices arises from the loop entropy effect (Skolnick, 1983) (the configurational entropy loss associated with constraining two ends of a random coil). In the case of coiled coils, loop entropy eliminates random coiled residues in between interacting helical stretches (Skolnick, 1983). Thus, coiled coils form a single interacting helical stretch, that can be preceded or followed by random coil residues. The later condition occurs for the 11−33 fragment where the short helical stretch prior to hydrophilic Asn is not strong enough to hold Asn in a helical conformation. On the contrary, the entire stretch of 10 residues prefers (due to the entropic reasons) to be disordered rather than to form an interacting helical stretch, followed by an internal random coil bubble. This situation is in contrast to that in the wild type and in fragment 8−30 where the stabilization of coiled coil structure starts (and ends) at the second (second to last) residue in the helical interface. This indicates that, in order to stabilize a unfavorable Asn residue in the helical interface, at least two consecutive hydrophobic residues are required (as in the case 8−30 fragment) to the left and to the right side of this residue.

Specificity of Coiled Coils. Equilibria in Fos-Jun system. In order to investigate a coiled coil system having the possibility of forming more than one structure, we chose the Fos-Jun transcriptional activator system. The experimental system itself consists of a unimolar mixture of Fos and Jun peptides at different concentrations (O'Shea *et al.*, 1989), where there is a possibility of formation of homodimers of both Fos and Jun as well as heterodimers of Fos with Jun. Jun homodimers are relatively stable, whereas Fos homodimers are of lower stability (O'Shea *et al.*, 1989) (see below). Experimentally, in the system containing a unimolar mixture of Jun and Fos, only heterodimers Jun-Fos are observed (O'Shea *et al.*, 1989).

Table 8: Predicted Dominant Species and Thermodynamical Parameters for 41 Residue Fos Sequence without CGG Linker

dominant species[a]		concn dependence of monomer/dimer ratio[b]		free energies	translational entropy loss[f]		
theory	expt	43 μM	1 mM		2 μM	43 μM	1 mM
1	1, 2	99:11	89:11	−165.7[c] −336.8[d] −5.4[e]	13.6	10.6	7.4

[a] The dominant species assigned for the entire concentration regime (43 μM−1 mM). "2" indicates the presence of dimers only, whereas "1" the presence of monomers. [b] The predicted ratio of the monomers to dimers calculated for different concentrations. [c] The configurational free energy of the unfolded state monomers in *kT*. [d] The configurational free energy of the folded state (coiled coil dimer) in *kT*. [e] The configurational free energy change upon folding (the free energy of dimer minus twice the monomer free energy) in *kT* units. [f] The translational entropy loss upon formation of the dimer multiplied by the reduced temperature for different concentrations (2 μM, 43 μM, 1 mM) in *kT*.

First, we performed the investigation of the stability of dimeric Fos-41 species reported to be minimally stable (O'Shea *et al.*, 1989) and used our method to estimate the stability of Fos coiled coils. The Fos-41 sequence corresponds to 41 residues, 160−200, from c-Fos oncoprotein with an additional H200Y mutation (O'Shea *et al.*, 1989; Schuermann *et al.*, 1991). Table 8 presents the fraction of monomeric and dimeric chains at various concentrations for Fos species without the CGG linker. Experimentally, these shorter Fos leucine zippers form dimers only at high concentration (O'Shea *et al.*, 1989), and their formation is concentration dependent. In Table 8, the free energies of unfolded monomers and dimeric coiled coils are given. Our prediction is too coarse to quantitate the concentration dependence of species; however, we do see a noticeable increase in the dimer population upon increasing the concentration.

We next present the results of our calculations for an equimolar mixture of leucine zippers from Fos and Jun oncoproteins. Because the experiment was done under strong renaturating conditions (O'Shea *et al.*, 1989) and the dimers were cross-linked (we assume that the only effect of cross-linking is to constrain chains), the only species present in the system are dimers, and the reaction under consideration is:

$$\text{Fos-Fos} + \text{Jun-Jun} \leftrightarrow 2\text{Fos-Jun} \qquad (13)$$

All of the peptides have CGG linkers at their N-terminal ends, as described in the experimental study (O'Shea *et al.*, 1989). In our statistical potential, the Cys−Cys interaction is the strongest. Even with no additional potential for cross-link formation, this interaction is sufficient to keep the two Cys residues interacting at all times (see the low values of accessible volume V_0 in Table 9). We use the approach described in the Method section to obtain the relevant configurational free energies of homodimeric Fos and Jun and heterodimeric Jun-Fos (note that heterodimer is preferred by symmetry—see eq 7a).

Table 9: Thermodynamical and Structural Parameters for the Fos-Fos, Jun-Jun and Fos-Jun Dimers with CGG Linkers

protein	dominant species P^a	dominant species E^b	rel ratio of the dimers[c]	free energies in coiled coil structure[d]	free energy of unfolded chains	vol occupied by the first bead V_0[g]	RMS in Å[h]
Fos-Fos			1.5	−371	−175.9[e]	107.6	1.7
					−19.6[f]		(1.3)
Jun-Jun			1.5	−389	−180.6[e]	144.3	1.0
					−27.5[f]		(0.3)
Fos-Jun	X	X	97	−384	−178.3[e]	107.1	3.4
					−27.7[f]		(1.1)

[a] The predicted (calculated) dominant species; "X" indicates which dimers are assigned. [b] The dominant species assigned from the experiment. [c] The fractions of each dimer. [d] The configurational free energy of the dimeric coiled coils in *kT*. [e] The configurational free energy of the unfolded, monomeric states in *kT*. [f] The configurational free energy change upon folding (the free energy of dimer minus twice the free energy of a monomer) in *kT* units. [g] Volume occupied by the first bead V_0 in Å^3 of the second chain (provided that the first bead of the first chain is pinned) and averaged over all simulations. [h] The average C_α RMS deviation of a single structure from the average structure (standard deviation is shown in parentheses).

Table 9 shows the predicted dominant species for the unimolar mixture of Fos (corresponding to residues 161−200 of c-Fos (Schuermann *et al.*, 1991) with an additional H200Y mutation) and Jun (corresponding to residues 279−318 of c-Jun (Schuermann *et al.*, 1991) with additional H318Y mutation) with CGG linkers at N-termini. The free energy of each sequence in the coiled coil dimer and average RMS deviations around the average structures obtained from MC simulations are also presented in Table 9. Only heterodimers of Jun-Fos are predicted to be present. This is consistent with the experimental observation (O'Shea *et al.*, 1989). Based on Table 9, homodimers of Jun have the highest relative stability, whereas Fos homodimers have the lowest relative stability. Because the average free energy of homodimers (half of the sum of the free energies of Jun and Fos) is higher than the free energy of heterodimers, heterodimers are preferred. The specificity of heterodimer formation can be accounted for by the relatively low stability of Fos-Fos homodimers. This observation is in agreement with the previous suggestion based on experiment that heterodimers are formed because of the low stability of Fos homodimers (O'Shea *et al.*, 1989). Thus, in this example the most stable species (Jun-Jun) is not dominant, and the preference for heterodimer formation comes from the instability of another component of the system (Fos homodimers).

The plots of the energy per residue presented in Figure 5 indicate that the reasons for the relatively low stability of Fos homodimers with respect to the Fos-Jun heterodimer are quite complex. Based on our simulations, the residues that contribute to the relative destabilization of the homodimeric Fos structure through interhelical interactions are located in region 5−13 (in particular, residues in positions 5a and 8d) and residues at positions 22d, 30e, and 33a. Position 8d (Leu in both Fos and Jun) is destabilized in Fos due to the interaction with two Thr residues (5a, 12a), whereas in Fos-Jun, Leu 8d interacts favorably with Ile 5a and Val 12a from Jun monomer. Positions 22d (Leu in both Fos and Jun) and 30e (Leu in Fos and Arg in Jun) are relatively destabilized in Fos-Fos homodimer due to unfavorable packing interactions and partial exposure to solvent. Position 33a (Lys in Fos and Val in Jun) is destabilized in Fos-Fos due to the unfavorable interactions with the corresponding 33a from the second chain. In general, the determinants of lower relative stability of Fos-Fos are related to both inter- and intrahelical interactions.

It is interesting to compare the absolute stabilities of dimers with GGC linkers. Table 9 shows the configurational free energy differences (free energy of folded structures minus twice the free energy of unfolded monomers) for both homodimers and heterodimer. The absolute stabilities are consistent with the relative stabilities. However, Jun-Fos heterodimer is now roughly of the same stability as Jun-Jun homodimer.

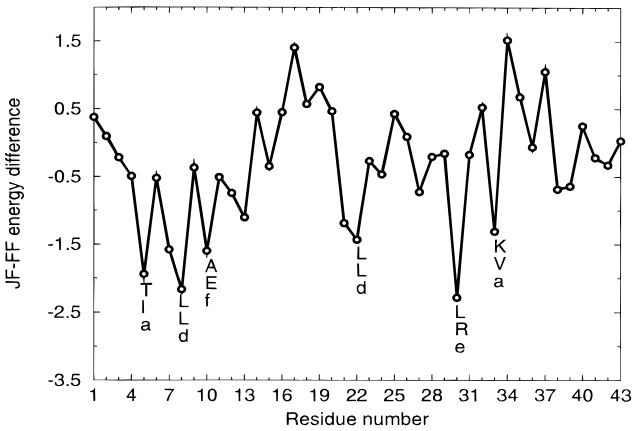

FIGURE 5: Plot of the energy per residue difference between Jun-Fos and Fos-Fos homodimers (computed as the energy per residue in Jun-Fos minus the energy per residue in Fos-Fos). The labels for residue numbers indicate (from the top) the Fos residue at a given position, the Jun residue, and the heptad positions, respectively. The largest, energetic relative destabilization of Fos-Fos homodimer appears to come from residues 4−13, as well as residues 22, 30, and 33. Most of those residues occupy positions at the helical interface (a, d, e, g).

DISCUSSION

In this paper, we have presented a method based on a transfer matrix treatment to calculate the monomer−dimer equilibrium in leucine zipper systems. For all cases examined, the prediction of dominant species is in good agreement with experiment. A detailed analysis of the simulations showed for some fragments of the GCN4 leucine zipper that quite substantial rearrangement can occur with respect to the wild type GCN4. Sometimes, as in the 11−33 fragment, many residues rearrange because of lack of stabilization of the coiled coil structure which when combined with loop entropy favors helix dissolution. In other cases (8−30), side chains slightly rearrange to minimize the conformational free energy. Thus, because fragments can locally adjust their conformation, one cannot simply assess subdomain stability based on the wild type data alone. Our results confirm the observation of Lumb *et al.* (1994) that some subdomains (fragments) of the GCN4 can form stable dimeric coiled

coils. On the basis of our data, we speculate that, in order to stabilize a hydrophilic Asn residue in a helical conformation in the hydrophobic core, it is necessary to have two sufficiently stable, interacting helical turns at both sides of this residue. This observation is an illustration of loop entropy, which in coiled coils prohibits randomly coiled residues between interacting helical stretches. If one of those helical turns is not strong enough to force a hydrophilic residue in the helical interface to be helical, then the entire fragment will be disordered.

In Fos-Jun system, heterodimers are formed because of the relative instability of Fos homodimers. This relative instability is predicted to come mainly from the region comprising residues 5−13 (interfacial Thr 5a, Leu 8d). Other interfacial residues (Leu 22d and Leu 30e, Lys 33a) also destabilize the Fos-Fos homodimer. The source of the relative instability comes from interhelical interactions, local packing, and different burial preferences. Our findings point out that the relative instability of Fos homodimers comes mainly from those residues occupying positions a and d of the heptad repeat. Residues at e and g positions seem to be less important. In general, our predictions are in agreement with experiment (Schuermann *et al.*, 1989, 1991; O'Shea *et al.*, 1992); however, further investigation is required to determine the specific role of any given residue in the preferential heterodimer formation. Because preferential heterodimer formation results from the relative instability of Fos homodimers, all of the residues responsible for the relative instability of Fos homodimers ultimately drive the process of heterodimerization.

Coiled coils are highly cooperative systems, and many of the phenomena observed in these systems can be explained by either cooperative or nonadditive effects. By cooperativity, we mean the additional energy a system gains if two or more events occur at the same time. The stabilization of Asn 16 in the helical interface in the wild type was explained by the necessity of maintaining the cooperative network of hydrogen bonds and by the stabilizing role of the cooperative pairwise interactions (Vieth *et al.*, 1994a, 1995a). Similarly, the effect of a single point N16V mutation was rationalized by the cooperative pairwise interactions and nonlocal compensation effects (Vieth *et al.*, 1995). The difference between the energetic profiles of the wild type and fragments can also be explained by smaller or larger structural rearrangements. The lack of stabilization of the hydrophilic Asn in the interface of the 11−33 subdomain can be explained by the insufficient cooperativity of the hydrogen bond network and side chain interactions in the region adjacent to this residue. Finally, the stabilization of Lys residues in the helical interface of Fos-Fos and Jun-Fos can also be explained by the stabilization resulting from the cooperative hydrogen bond network as well as the cooperative packing interactions.

The method we presented in this paper is, in principle, general enough to calculate the free energy of folding of any small protein assuming that the final structure is known or can be deduced. Thus, the effect of the mutations could be studied for both folded and unfolded states and compared to experimental data. In principle, the method presented here is not limited to lattice models but could be applied to any protein model that has a small number of local minima for the corresponding set of peptide fragments. Work in this direction is now in progress.

The work presented in this paper sets the stage for the lattice free energy calculations of proteins. It also builds a groundwork for fast coiled coil prediction algorithms. In particular, this work provides the basis for the automated assessment of the heterodimerization ability of leucine zippers that is of great importance in studies of transcriptional regulation. While coiled coil systems have been recognized as the simplest examples of proteins and the methodology presented here has, even for these systems, limitations and problems, this paper provides encouragement for further theoretical studies along these lines of proteins and biological systems in general.

ACKNOWLEDGMENT

Helpful discussions with Prof. Charles L. Brooks, III, are acknowledged. We thank the referees for their useful suggestions. A.K. is an International Research Fellow of the Howard Hughes Medical Institute.

APPENDIX A

Description of the Lattice Model.

(1) Geometric Representation and Move Set. There are in principle 90 possible ways of connecting two consecutive α-carbons on this lattice, but some geometric restrictions for consecutive two and three sets of vector occurrences limit this number by roughly 60%. Crystal structures of proteins from the Brookhaven Protein Data Bank, PDB, can be represented by the lattice model with an average, root mean square deviation, RMS, of 0.6−0.7 Å (Godzik *et al.*, 1993). For the folded state calculations, the Monte Carlo move set consists of two bond moves, three bond rearrangements, small shifts of larger chain pieces, chain end modifications, and rotamer equilibration. One Monte Carlo cycle for this system is considered to have $(N - 2)M$ two bond moves, $2M$ two bond moves, M shifts of the chain pieces, and $M(N - 3)$ three bond moves, where M is the number of chains (in this calculation $M = 2$) and N is the number of α-carbons. Each simulation run consisted of $(200000)M$ cycles (Vieth *et al.*, 1995a).

(2) Interaction Scheme. The entire potential (with the exception of the hydrogen bond term) is based on a statistical analysis of a set of high resolution crystal structures from the PDB database. The use of statistical potential to estimate protein stability has been attempted previously. For example, Bryant and Lawrence showed that the frequency of occurrence of charged residues in the proteins from PDB obeys Coulombs law; however, the dielectric constant is too high (Bryant & Lawrence, 1991). Thus, such statistical potentials may provide some insight toward understanding protein stability. Based on the folding of the Hodges sequences (Hodges *et al.*, 1981) and a test of the dynamic stability of assembled dimers, the scaling factors for the different energy terms were chosen to keep the helix content of the noninteracting chains below 50%, as well as to maintain a proper balance of the short range and long range interactions. The scaling factors for all of the coiled coils systems studied below are the same as in the previous study on the GCN4 leucine zipper folding from random chains (Vieth *et al.*, 1994b) as well as in a previous study of oligomeric equilibria (Vieth *et al.*, 1995a). A detailed description of all energy terms is presented elsewhere (Vieth *et al.*, 1995a).

The total energy of the entire system is given by (Vieth *et al.*, 1995a):

$$E_{tot} = E_{short} + E_{long} = E_{HB} + E_{\beta} + 0.25E_{14} + 0.5E_{rot} + 0.5E_{one} + 5E_{pair} + 4.25E_{tem} \quad \text{(A1)}$$

where E_{tot} is the total energy of the system, $E_{short(long)}$ is the short (long) range interaction energy. E_{HB} is the hydrogen bond potential, E_{β} is the $C_{\alpha}-C_{\beta}$ short range orientational potential, E_{14} is the potential associated with three consecutive $C_{\alpha}-C_{\alpha}$ vectors, E_{one} is the one body term, E_{pair} is the pair potential, and E_{tem} is the cooperative pair potential. All parameters are available via anonymous ftp (Vieth *et al.*, 1994b).

For all of the simulations in the present work, the reduced temperature, T_{red} (used to determine acceptance ratio of the moves via a standard asymmetric Metropolis scheme (Metropolis *et al.*, 1953)), was set to 1.85. This temperature was chosen because, in the original folding simulations of GCN4, this corresponded to native conditions (Vieth *et al.*, 1994b).

APPENDIX B

Description of the Transfer Matrix Treatment for the Unfolded State. For a chain of length N, we define the configurational partition function (rotations are included because we consider all possible orientations of the first two vectors) as the sum of the statistical weights for all of the possible conformations of $N - 1$ bonds (Skolnick & Kolinski, 1992):

$$Z^{N}_{int,90} = \sum_{i=1}^{N_{states}} e^{-\beta E_i} = \sum_{i,j,k,l=1}^{90} \left(\prod_{n=1}^{N-4} \mathbf{U}^{n}_{ijkl}\right) \quad \text{(B1)}$$

where $\mathbf{U}^{n}_{ijkl}$ is the $90 \times 90 \times 90 \times 90$ statistical weight matrix for the n-th fragment. A single element of this matrix is defined as (i, j, k, l represent four vectors for each fragment):

$$\mathbf{U}^{n}_{ijkl} = \delta_{ijk}\delta_{jkl} \exp(-\beta(E^{14}_{ijkl} + E^{HB}_{ijk} + E^{one}_{i} + E^{HM}_{ijkl} + E^{ter}_{ijkl}) \sum_{i1}^{N_i} \exp(-\beta E^{rot}_{i1}) \quad \text{(B2)}$$

where δ_{ijk} is defined by analogy to an ortho-normal basis set as follows: $\delta_{ijk} = 1$ if three consecutive vectors i, j, k are allowed, and $\delta_{ijk} = 0$ otherwise (due to the excluded volume and geometric consideration). The treatment is schematically shown in Figure 6. Let us note that eq B2 would provide the exact number of accessible states of the chain if all of the energy terms were equal to zero. Multiplication of 4 element matrices leading to the statistical weight matrix associated with extension of a chain by one bond is done as follows (see also Figure 6):

$$\mathbf{U}^{new}_{ijkl} = \sum_{m=1}^{90} \mathbf{U}^{n-1}_{imjk}\mathbf{U}^{n}_{mjkl} \quad \text{(B3)}$$

where n denotes the segment number and the summation is done over 90 possible orientations of the second vector of the $(n - 1)$-th segment that is by construction the same as first vector of the n-th segment. The resulting matrix elements have statistical weights associated with the specific chain conformation having the first vector i and three last vectors j, k, and l and all possible combinations of vectors in between.

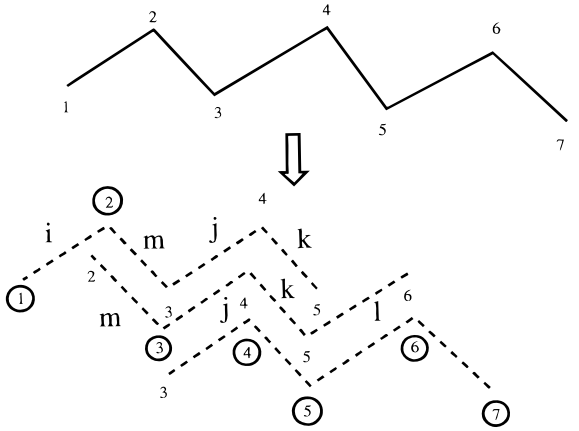

FIGURE 6: Schematic representation of internal free energy calculation for the seven residue chain in the unfolded state. A seven residue segment is divided into three four bond (or five residue) segments. For each of them statistical weight matrices are calculated in all 90^4 possible conformations. For the first segment, interactions of residues 1 and 2 with everything else are calculated. For the middle segment, interactions of the second residue are computed (here, residue 3), and for the last fragment interactions of the last four residues are evaluated. Thus, no interactions are double counted. This picture also shows the consecutive vectors *i, m, j, k, l* that would be used to construct the statistical weight of the matrix element *i, j, k, l*. This element would consist of a sum of 90 weights over all possible conformations of the *m*-th vector.

APPENDIX C

Monte Carlo Simulations of the Folded State. The main purpose of the simulations is to calculate the average energy E_0, the probability of the average energy $P(E_0)$, and degeneracy of the average energy level, $n(E_0)$. First, we need to generate starting structures in the conformation of dimeric coiled coil. The protocol for generation of the starting structures is described in Figure 7. Then long, unrestrained Monte Carlo simulations (200 000 MC cycles) are run to obtain the sampling in the neighborhood of the dimeric coiled coil structures for each sequence as described previously (Vieth *et al.*, 1995a). The points from the energy plateau regions for each simulation were used to construct histograms of the energy distributions. Energy bins of $1kT$ width were used. Since the histograms were collected at the same temperature, we used constant temperature WHAM equations (Kumar *et al.*, 1992) to obtain the final energy probability distribution for each system under consideration (different sequences in the dimeric coiled coil structure). From the final histogram (Figure 8), the average energy of the system was computed as well as the probability of a system being in the average energy level.

Monte Carlo simulations of the coiled coil state were performed to estimate the degeneracy of the average energy level E_0. In this case, we collected 2000 structures with the energy within $4kT$ from the previously computed average energy. The collection of the structures is done every 50 Monte Carlo cycles. Because the simulation runs are relatively short and we collect only 2000 structures, not all of the possible structures around the folded state are sampled. To enrich the sampling, we assume that the entire ensemble of folded conformations can be estimated as a product of

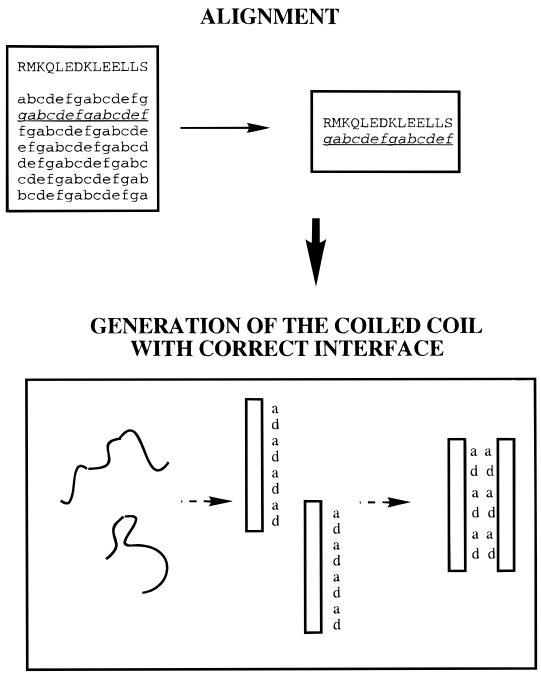

FIGURE 7: Schematic representation of the alignment of a sequence to the heptad repeat and generation of the double stranded parallel coiled coil geometry for the aligned sequence. Alignment of a sequence to the periodic heptad repeat can start from any of the seven letters (a−g). Each of the seven alignments has its own score for a given sequence. The scoring function consists of a pair potential, assuming an idealized interchain coiled coil contact map (i.e., a_i interacts with a_i, g_{i-1}, d_i, d_{i-1}), a one body burial term (everything except residues in a or d positions is unburied), and an idealized hydrophobic moment energy. The positions of the atoms required to compute all energetic contributions were taken from the crystal structure of the GCN4 leucine zipper. Alignment for the parallel, homodimeric structures can be viewed as a simultaneous rotation around the 7-fold axis for each helix. Then, the alignment with the best score is selected. In the figure, the underlined alignment that starts from letter d is selected. For all of the tested sequences, the alignment is consistent with that based on the Lupas scoring function (Lupas *et al.*, 1991) as well as with the experimental crystal structures (O'Shea *et al.*, 1991). Having chosen the best alignment for each sequence, two helices are built on the lattice, translated to a neighborhood of one another with a and d residues in the interface. The interhelical contact restraints between corresponding a and d residues together with helical biases for the secondary structure are then applied, and the structure is equilibrated by short Monte Carlo simulations.

overlapping three residue backbone segments occurring in the simulation. In addition, all of the conformations generated in that manner are treated as having the same energy (the average energy around which sampling is performed). Thus we calculate the number of states per each chain independently, with no explicit interchain correlation. These assumptions probably slightly overestimate the number of structures having the average energy. With these assumptions, $n(E_0)$ is given by:

$$n(E_0) = \prod_{\gamma=1}^{2} \sum_{i,j,k}^{90} \left(\prod_{n=1}^{N-3} \mathbf{F}_{ijk}^{n,\gamma} \right) \tag{C1}$$

from the combination of overlapping three vector segments occurring in the simulation. The **F** are $90 \times 90 \times 90$ transfer matrices for each of the $N - 3$ segments in chain γ. The elements of each matrix are defined as:

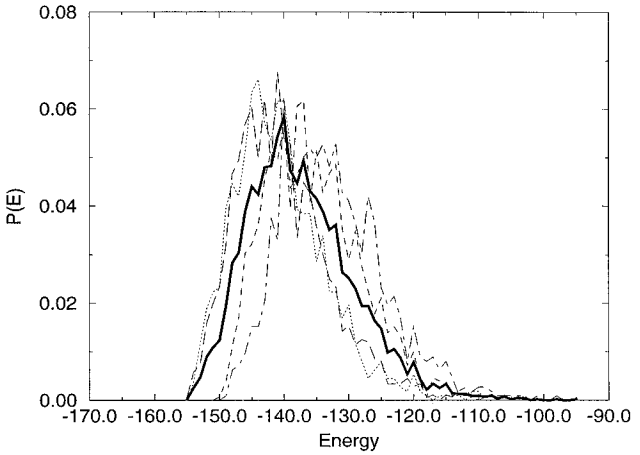

FIGURE 8: Energy probability distribution for one of the simulations of the 8−30 fragment. The dashed lines represent the probability distribution from four independent simulations. The bold line is obtained by the WHAM method (Kumar *et al.*, 1992) from these four independent simulations.

$$\mathbf{F}_{\mathrm{ijk}}^{n} = \omega_{ijk} \sum_{i1}^{N_{\mathrm{ij}}} \omega_{i1}^{ij} \sum_{i2}^{N_{jki}} \omega_{i2}^{jk} \tag{C2}$$

where the backbone weights ω_{ijk} are equal to unity (zero) if three consecutive vectors i, j, k (do not) occur during the simulation. ω_{i1}^{ij} equals unity (zero) if a given rotamer $i1$ occurs (does not occur) in the course of the Monte Carlo simulation for a given backbone conformation ij. N_{ij} indicates the maximum number of rotamers of a given type for a given backbone conformation i, j.

The choice to estimate the number of folded conformations having the average energy by construction from three bond fragments is a compromise between lattice model restrictions and sampling efficiency. First, we have three vector restrictions in our lattice model (no longer range restrictions). Second, the statistics for occurrence in the simulation of three vector fragments is acceptable—the collection of 4000 instead of 2000 structures gives practically the same $kT \ln(n(E_0))$ (within $1kT$). Using two vector segments would generate some states that are not permitted by some three vector restrictions. In contrast, there are not enough statistics for four vector states.

(2) Configurational Partition Function. Substituting eqs C1 and C2 into eq 11, we get the expression for the partition function of the folded state:

$$Z_{\mathrm{conf,D}} = N_{\mathrm{D1}} N_{\mathrm{D2}} \exp(-E_0/kT) \prod_{\gamma=1}^{2} \prod_{i,j,k}^{90} \left(\prod_{n=1}^{N-3} \mathbf{F}_{ijk}^{n,\gamma} \right) / P(E_0) \tag{C3}$$

Let us note that a similar treatment can be done for any lattice model of a protein. The entropy associated with energetic fluctuations around the average energy level (see eq 12) is usually on the order of $3kT$.

The difference between the local volume factorization method (Vieth *et al.*, 1995a,b) previously used in the determination of entropy for the equilibria between dimers, trimers, and tetramers for mutants of GCN4 leucine zipper with the transfer matrix treatment has been examined. Both methods predict the same dominant species for 6 out of 8 cases (except VL and IL mutants), and both agree with experiment in 5 out of 8 cases. Both methods show a similar

trend in the entropy favoring lower order oligomers; the entropy in the transfer matrix treatment favors dimers over trimers (trimers over tetramers) on average by 3.2*kT*/monomer (3.1*kT*/monomer), and in the local volume treatment by 1.5*kT*/monomer (1.3*kT*/monomer). While both methods are approximate, the transfer matrix treatment seems to be a more natural choice for the lattice models. It allows for the treatment of the unfolded chains and also includes explicit short range correlations. The transfer matrix treatment gives roughly a 2 times larger number of states per residue than the local volume factorization approximation due to the fact that for one "volume" state there can be multiple lattice vector states occupying this volume.

REFERENCES

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, J. E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535−542.

Bryant, S. H., & Lawrence, C. E. (1991) *Proteins 9*, 108−119.

Cohen, C., & Parry, A. H. (1990) *Proteins 7*, 1−15.

Cohen, C., & Parry, D. A. D. (1986) *Trends Biochem. Sci. 11*, 245−248.

Crick, F. H. C. (1952) *Nature 170*, 882−882.

Crick, F. H. C. (1953) *Acta Crystallogr. 6*, 689−697.

DeLano, W. L., & Brunger, A. T. (1994) *Proteins 20*, 105−123.

Flory, P. (1969) *Statistical Mechanics of Chain Molecules*, Wiley, New York.

Fraser, R. D. B., & MacRae, T. P. (1971) *Nature 233*, 138−140.

Frauenfelder, H., Parak, F., & Young, R. D. (1988) *Annu. Rev. Biophys. Biophys. Chem. 17*, 451−479.

Godzik, A., Kolinski, A., & Skolnick, J. (1993) *J. Comput. Chem. 14*, 1194−1202.

Godzik, A., Kolinski, A., & Skolnick, J. (1995) *Protein Sci. 4*, 2107−2117.

Hai, T., Fang, L., Coukos, W. J., & Green, M. R. (1989) *Genes Dev. 3*, 2083−2090.

Harbury, P. B., Zhang, T. Kim, P. S., & Alber, T. (1993) *Science 262*, 1401−1407.

Harrison, S. (1991) *Nature 353*, 715−719.

Herschbach, D. R. (1959) *J. Chem. Phys. 31*, 1652−1661.

Hodges, R. S., Saund, A. S., Chong, P. C. S., St-Pierre, S. A., & Reid, R. E. (1981) *J. Biol. Chem. 256*, 1214−1224.

Holtzer, A. (1995) *Biopolymers 35*, 595−602.

Johnson, P., & Smillie, L. B. (1975) *Biochem. Biophys. Res. Commun. 64*, 1316−1322.

Kolinski, A., & Skolnick, J. (1994) *Proteins 18*, 338−352.

Krystek, S. R., Jr., Bruccoleri, R. E., & Novotny, J. (1991) *Int. J. Peptie Protein Res. 38*, 229−236.

Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., & Rosenberg, J. M. (1992) *J. Comput. Chem. 13*, 1011−1021.

Landshultz, W. H., Johnson, P. F., & McKnight, S. L. (1988) *Science 240*, 1759−1764.

Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181−293.

Lifson, S., & Roig, A. (1961) *J. Chem. Phys. 34*, 1963−1974.

Lovejoy, B., Seunghyon, C., Cascio, D., McRorie, D. K., DeGrado, W. F., & Eisenberg, D. (1993) *Science 259*, 1288−1293.

Lumb, K. J., Carr, C. M., & Kim, P. S. (1994) *Biochemistry 33*, 7361−7367.

Lupas, A., Van Dyke, M., & Stock, J. (1991) *Science 252*, 1162−1164.

Mayer, J. E., & Mayer, M. G. (1963) *Statistical Mechanics*, Wiley: New York.

McLachlan, A. D., & Stewart, M. (1975) *J. Mol. Biol. 98*, 293−304.

McQuarrie, D. A. (1976) *Statistical Mechanics*, Harper & Row, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953) *J. Chem. Phys. 21*, 1087−1092.

Nilges, M., & Brunger, A. T. (1993) *Proteins 15*, 133−146.

O'Shea, E. K., Rutkowski, R., Stafford, W. F., III, & Kim, P. S. (1989) *Science 245*, 646−648.

O'Shea, E. K., Klemm, J. D., Kim, P. S., & Alber, T. (1991) *Science 254*, 539−544.

O'Shea, E. K., Rutkowski, R., & Kim, P. S. (1992) *Cell 68*, 699−708.

O'Shea, E. K., Lumb, K. J., & Kim, P. S. (1993) *Curr. Biol. 3*, 658−667.

Perutz, M. (1992) *Protein structure. New approaches to disease and therapy*, W. H. Freeman, New York.

Phillips, G. N. J., Fillers, J. P., & Cohen, C. (1986) *J. Mol. Biol. 192*, 111−131.

Poland, D., & Scheraga, H. A. (1970) *Theory of helix−coil transitions in biopolymers*, Academic Press, New York.

Privalov, P. E. (1992) Physical basis for the stability of the folded state, in *Protein folding*, pp 83−126, W. H. Freeman, New York.

Ptitsyn, O. B. (1987) *J. Protein Chem. 6*, 273−293.

Schuermann, M., Neuberg, M., Hunter, J. B., Jenuwein, T., Ryseck, R. P., Bravo, R., & Muller, R. (1989) *Cell 56*, 507−516.

Schuermann, M., Hunter, J. B., Hennig, G., & Muller, R. (1991) *Nucleic Acids Res. 19*, 739−746.

Skolnick, J. (1983) *Macromolecules 16*, 1069.

Skolnick, J., & Helfand, E. (1980) *J. Chem. Phys. 72*, 489.

Skolnick, J., & Kolinski, A. (1992) *J. Mol. Biol. 223*, 499−531.

Smeal, T., Angel, P., Meek, J., & Karin, M. (1989) *Genes Dev. 3*, 2091−2100.

Vieth, M., Kolinski, A., Brooks, C. L. I., & Skolnick, J. (1994a) *J. Mol. Biol. 237*, 361−367.

Vieth, M., Kolinski, A., & Skolnick, J. (1995b) *J. Chem. Phys. 102*, 6189−6193.

Vieth, M., Kolinski, A., Brooks, C. L., III, & Skolnick, J. (1995a) *J. Mol. Biol. 251*, 448−467.

Vieth, M., Skolnick, J., Kolinski, A., & Godzik, A. (1994b) *Available via anonymous ftp from ftp.scripps.scripps.edu in pub/(MCDP.info, MCDP.Z).*

Zhang, L., & Hermans, J. (1993) *Proteins 16*, 384−392.

Zimm, B. H., & Bragg, J. K. (1959) *J. Chem. Phys. 31*, 526−535.

BI9520702