Gordon M. Crippen
College of Pharmacy,
University of Michigan,
Ann Arbor, MI 48109-1065

Received 24 March 2004;
accepted 17 June 2004

Statistical Mechanics of Protein Folding by Cluster Distance Geometry

Published online 31 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20118

Abstract: *This is our second type of model for protein folding where the configurational parameters and the effective potential energy function are chosen in such a way that all conformations are described and the canonical partition function can be evaluated analytically. Structure is described in terms of distances between pairs of sequentially contiguous blocks of eight residues, and all possible conformations are grouped into 71 subsets in terms of bounds on these distances. The energy is taken to be a sum of pairwise interactions between such blocks. The 210 energy parameters were adjusted so that the native folds of 32 small proteins are favored in free energy over the denatured state. We then found 146 proteins having negligible sequence similarity to any of the training proteins, yet the free energy of the respective correct native states were favored over the denatured state.* © 2004 Wiley Periodicals, Inc. Biopolymers 75: 278–289, 2004

Keywords: *thermal denaturation; globular proteins; canonical partition function; conformational sampling; distance geometry*

INTRODUCTION

Computational approaches to modeling protein folding generally involve extensive molecular dynamics[1–4] or Monte Carlo[5–9] simulations. These methods have been shown to work for small systems, and in principle they eventually will produce a Boltzmann distribution of configurations for even solvated proteins, but there remain lingering concerns about the adequacy of sampling for any finite calculation.[10] Recently, we have been exploring an entirely new approach called SMEUSE (*S*tatistical *M*echanics *E*nabled *U*sing *S*eparable *E*nergies). Here, the central idea is that all the possible configurations of the system must be described in terms of some set of parameters that permit the construction of an energy-like function consisting of a sum of terms, each of which depends on a separate, small subset of the configurational parameters. Then the canonical partition function breaks up into a product of simple integrals over each subset of the parameters. Such a partition function is vastly easier to evaluate than integrating over all parameters simultaneously or approximating it by a stochastic process in the high-dimensional parameter space. The difficulty is in devising the set of configurational parameters and corresponding energy function, particularly when standard empirical energy functions, based on the customary classification of physical effects, are unsuitable for this approach.

In our first SMEUSE study,[11] we described protein conformations in terms of wavelet transforms of the C^{α} trace coordinates versus sequence position in the polypeptide chain. The associated energy function depended on conformation and the amino acid content of different segments of the chain. We were able to adjust its 26 parameters so that a training set of seven proteins favored the native over the denatured state.

Correspondence to: Gordon M. Crippen: email: gcrippen@umich.edu
Biopolymers, Vol. 75, 278–289 (2004)

Then another 480 unrelated test proteins also had lower free energies for the configuration space around their respective Protein Data Bank (PDB) native conformations compared to the denatured state. This remains a promising approach even though it is difficult to interpret the energy function in conventional terms.

Here we initiate an alternative SMEUSE parameterization for protein folding based on a recent generalization of distance geometry where the objects are not single geometric points, but sets of points.[12] Possible advantages are a more direct connection to customary ideas about through-space interactions between amino acid residues, more natural incorporation of known limitations on packing density, and eventual incorporation of additional constraints, such as disulfide crosslinks.

METHODS AND RESULTS

All calculations described in this section were carried out in MOE using the SVL computer language.[13]

Polypeptide Conformations in Terms of Interresidue Distances

In this work we have considered only small and simple proteins consisting of a single polypeptide chain involving only the standard 20 amino acid residue types, without crosslinks or substantial ligands. For reasons of computational convenience, attention has been restricted to chain lengths n no greater than 128 residues. As a first approximation, let the polypeptide chain be represented as only one point per residue, taken to be the C^{α} atom. Assuming standard bond lengths, bond angles, and fixed planar trans peptide bonds, the distance between successive C^{α} atoms is fixed, as in a freely jointed chain model, although even allowing all possible ϕ and ψ dihedral angles in each residue is more restrictive than the freely jointed chain.

In order to describe conformations of such chains independent of overall translations and rigid proper rotations, a natural choice of parameters is the set of interresidue distances, $\{d_{ij},\ i,j = 1, \ldots, n\}$, but there are $n(n-1)/2$ of these, whereas a freely jointed chain in three dimensional space would have $3n - 6 - (n-1) = 2n - 5$ net degrees of freedom, subtracting three translational and three rotational degrees of freedom and the $n-1$ fixed $d_{i,i+1}$ distance constraints. Indeed, short polypeptide chains having freely variable ϕ and ψ dihedral angles do exhibit $2n - 5$ degrees of freedom. Consider tetraalanine, Ala_4, in C^{α} representation where 5000 random $\phi\psi$ conformations were sampled and displayed as dots in Figure 1 scattered in a coordinate system consisting of the $2n - 5 = 3$ variable distances, $d_{1,3}$, $d_{2,4}$, and $d_{1,4}$. If there are m degrees of freedom, then the number of conformations c within radius r of the center of the cloud should vary as r^m, and in fact such a log–log plot of c vs. r agrees with $m = 3$ for small r (plot not shown). While it is certainly possible for the allowed conformations of cyclic molecules to occupy complicated lower-dimensional subspaces,[14] the conformation spaces of polypeptide chains have locally $2n - 5$ dimensions.

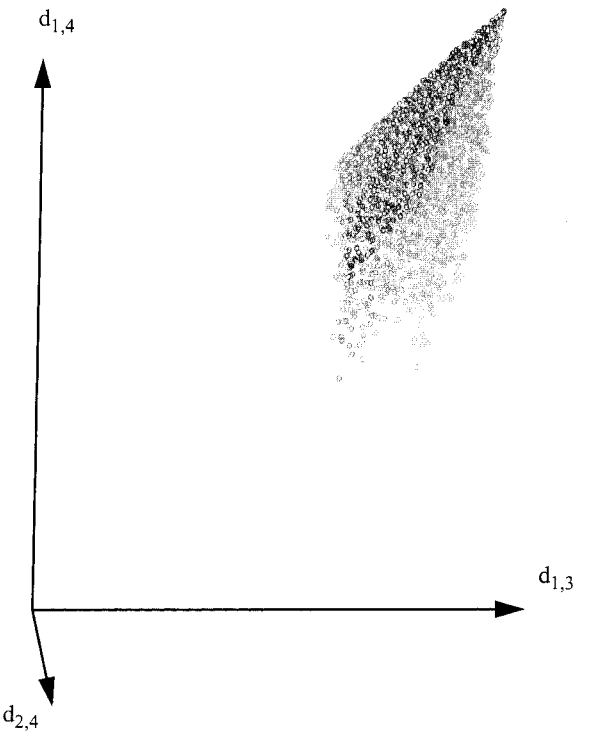

FIGURE 1 The conformation space for Ala_4 in terms of the three nontrivial inter-C^{α} distances. Note how the lower bounds of $d_{1,3}$ and $d_{2,4}$ necessarily increase as $d_{1,4}$ approaches its upper bound.

Globally speaking, Figure 1 clearly shows additional conformational restrictions. Each of the three variable interresidue distances has upper and lower bounds due to chain connectivity, flexibility, and steric hindrance. If they were otherwise independent, the figure should show a simple rectangular solid. Note, however, that as $d_{1,4}$ approaches its upper bound (i.e., the chain becomes nearly fully extended), the lower bounds of $d_{1,3}$ and $d_{2,4}$ also increase, giving rise to the point at the top of the scattering of conformers. At the other extreme of small $d_{1,4}$, the allowed values of $d_{1,3}$ and $d_{2,4}$ become negatively correlated, as shown in the cross-section view in Figure 2. Such effects are readily understood intuitively and can be derived from the fixed $d_{i,i+1}$ values plus the triangle inequality, $d_{i,k} \leq d_{i,j} + d_{j,k}$.

The question, then, is how to describe the set of all possible conformations of a polypeptide chain in a compact and convenient form while at least approximately accounting for the true number of degrees of freedom and the rather subtle interrelations we can see in even a tetrapeptide. As in standard distance geometry calculations of conformations,[15] one can derive upper and lower bounds, $l_{ij} \leq d_{ij} \leq u_{ij}$, for each of the $n(n-1)/2$ distances, and all conformations must lie inside this initial hyperrectangle in $n(n-1)/2$ dimen-

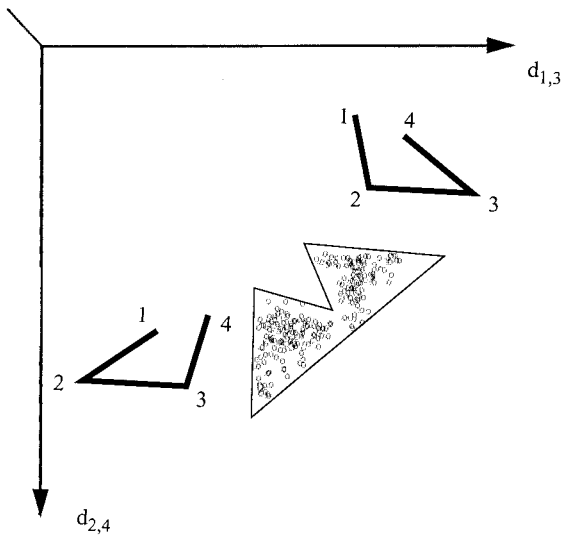

FIGURE 2 Detail of Figure 1 when $d_{1,4}$ is small, showing that the region of allowed conformations is no longer convex because $d_{1,3}$ and $d_{2,4}$ are negatively correlated.

sions, which we will refer to as the *a priori* "range." Figures 1 and 2 clearly show that the allowed conformations lie in a lower dimensional subspace, they do not completely fill such an initial rectangular range, and the volume to be described is not even necessarily convex. An approximate way to deal with this is to break up the initial range into smaller ranges that are mutually nonoverlapping but have contracted their bounds to better outline the volume of allowed conformations, as shown in Figure 3. The union of the ranges still accounts for all possible conformations, but the disallowed volume enclosed by them decreases. This approach has been used before to describe the conformations of small molecules,[16,17] but for even $n = 128$ residues, this is quite an undertaking in 8128 dimensions. Any sort of conformation space can be described to any desired accuracy as the union of many nonoverlapping ranges, but the number of ranges increases rapidly with the dimensionality and required accuracy.

In order to reduce the dimensionality of the task, we turned to a recent generalization of distance geometry called cluster distance geometry,[12] where the polypeptide chain is subdivided into blocks of b sequentially adjacent residues, and the variables are the sums of squared interresidue distances between blocks I and J.

$$D_{IJ} = \sum_{i \in I} \sum_{j \in J} d_{ij}^2 \tag{1}$$

Here we have used $b = 8$, so that for $n = 128$ there are 16 blocks, and ranges involve $16 \times 15/2 = 120$ dimensions. While this representation of the chain has obviously even lower resolution than one point per residue, considerable *a priori* information can be built in regarding limitations on polypeptide chain packing, and the D_{IJ} are easily interrelated at the level of triangle inequality reasoning.[12] In particular, the bounds $I_{IJ} \leq D_{IJ} \leq U_{IJ}$ that determine the initial overall range covering all polypeptide conformations are readily calculated from n and b, as described in Ref. 12.

Separable Energy Model

The guiding principle of SMEUSE is that corresponding to the conformational variables employed, there must be an energy-like function that consists of a sum of terms, each of which depends on a small disjoint subset of the variables. Here we take the terms to be simple square well potentials depending on the summed interblock distance, D_{IJ}, and the amino acid compositions of the two blocks without regard to sequence separation or radii of gyration of the blocks.

$$E_{tot} = \sum_{I<J} E(D_{IJ}, \boldsymbol{t}_{IJ}) \tag{2}$$

Equation (2) assumes that the contribution to the total energy from each pair of blocks is independent of the other D_{IJ} values, even though these quantities are interdependent in order to correspond to realizable conformations in three-dimensional space.[12] The assumption is that these relations are adequately approximated by the set of ranges used to describe all possible conformations, rather than considering only a single rectangular range.

The individual square well terms are very simple.

$$E(D_{IJ}, \boldsymbol{t}_{IJ}) = \begin{cases} \boldsymbol{a} \cdot \boldsymbol{t}_{IJ} & \text{for } D_{IJ} < \kappa \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

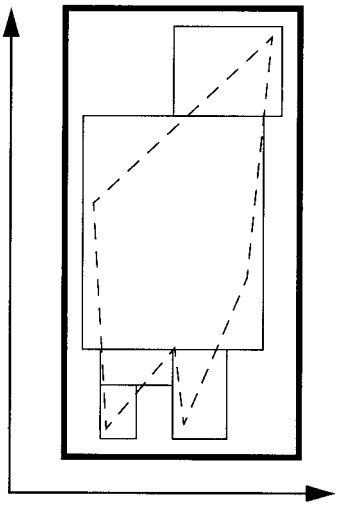

FIGURE 3 A two-dimensional example of how the region of allowed conformations in Figure 1 can be represented as the union of nonoverlapping rectangles. The *ab initio* bounds for all conformations are shown as the heavy rectangle (plus a small margin for clarity), the true conformation space is enclosed in dashed lines, and the five ranges are indicated by light rectangles.

The interblock sequence composition vector $\boldsymbol{t}_{IJ}$ has 210 components, each being the number of residue pairs from the two blocks having a particular (unordered) residue pair type, just as D_{IJ} consists of a sum over pairs of points between the two blocks in Eq. (1). Thus, the sum of the components of each $\boldsymbol{t}_{IJ}$ is always b^2, except when the blocks involve virtual residues, as noted below. Clearly this is a rough approximation to the real interactions in that any random permutation of the amino acid sequence within a block will give the same $\boldsymbol{t}_{IJ}$ vectors. Two blocks are viewed as interacting across all their constituent residue pairs as long as the overall D_{IJ} is below a fixed cutoff, κ, regardless of whether the blocks are spatially extended or compact. The interaction may be either positive (unfavorable) or negative (favorable), depending on the signs and magnitudes of the components of $\boldsymbol{a}$. Steric repulsion for very small values of D_{IJ} is treated as the responsibility of the set of ranges to enforce $L_{IJ} \leq D_{IJ}$ for some appropriate lower bounds, rather than building in an energetic penalty. Equations (2) and (3) should be viewed as a very simple initial energy model that can doubtless be substantially improved in future studies.

In this work, the only adjustable parameters in the energy function are the 210 components of the $\boldsymbol{a}$ vector and the cutoff κ. We use the same parameters for all terms in Eq. (2) without regard to total chain length n or sequence separation $|I - J|$. Presumably a different choice for $b = 8$ would require readjusted $\boldsymbol{a}$ and κ. How the parameters are adjusted is discussed below.

Conformational Ranges Suited to Proteins and the Energy Function

There are many imaginable schemes for subdividing the initial *ab initio* range into smaller ranges that more precisely describe the full set of possible conformations. The guiding principles used here are first that extended conformations seen in the denatured state of proteins need not be so precisely described as compact, native-like conformations. Second, there is no need to discriminate between even two compact conformations that are viewed as equivalent by the energy function. Consequently, the final set of ranges depends on the form of the energy function and on the set of proteins used for training it.

Since our model considers only a single polypeptide chain, the proteins in the training set should be those having native conformations stabilized strictly by noncovalent intrachain interactions. From the total Protein Data Bank[18] (PDB) of over 26,000 experimentally determined three-dimensional structures, PDB Select[19] is a subset where the proteins have sequences that differ by at least a small amount, according to a formula that permits a smaller fraction of identical residues for longer chains. Out of the 5416 entries in the PDB Select 90% list of April 2002, we found 96 X-ray crystal structures apparently involving only one, short polypeptide chain without substantial ligands, such as heme groups. Further scrutiny resulted in only 32 entries consisting of only a single polypeptide chain of length no greater than 128 residues that seem to fold as monomers under reasonably standard conditions to a compact structure having a radius of gyration no more than 30% greater than the minimum for the given chain length.[20] Furthermore, no pair of these 32 chains has greater than 90% sequence identity after optimal sequence alignment, and the root mean square deviation (rmsd) between matching aligned residues after the usual optimal rigid body superposition[21] is greater than 3 Å. As shown in Table I, the final training set involves considerable diversity of fold types.[22]

Table I Training Set Proteins

PDB Entry	No. Residues	Type[a]	Native Range	Z_{nat}/Z_{tot}
1AIX	106	β	70	0.497
1ACF	125	$\alpha + \beta$	56	0.498
1BK2	57	β	71	0.500
1BM8	99	$\alpha + \beta$	55	0.499
1BYW.A	110	$\alpha + \beta$	53	0.498
1C44.A	123	$\alpha + \beta$	61	0.499
1COA.I	64	$\alpha + \beta$	57	0.499
1CQY.A	99	β	63	0.500
1DHN	121	$\alpha + \beta$	69	0.498
1DT4.A	73	$\alpha + \beta$	67	0.499
1ENH	54	α	58	0.500
1EW4.A	106	$\alpha + \beta$	62	0.499
1G9O.A	91	Membrane[b]	63	0.500
1HEY	127	α/β	71	0.500
1I2T.A	61	α	54	0.499
1JWO.A	97	$\alpha + \beta$	71	0.499
1MIL	104	$\alpha + \beta$	60	0.496
1MJC	69	β	63	0.500
1OPS	64	β	71	0.499
1PGB	56	$\alpha + \beta$	52	0.500
1PHT	83	β	63	0.499
1PTF	87	$\alpha + \beta$	71	0.499
1QAU.A	112	β	59	0.499
1TEN	89	β	66	0.498
1TMY	118	α/β	71	0.499
1TUL	102	β	65	0.498
1UBI	76	$\alpha + \beta$	71	0.983
1VCC	77	$\alpha + \beta$	71	1.00
1WHI	122	β	53	0.499
2IGD	61	$\alpha + \beta$	68	0.499
3IL8	68	$\alpha + \beta$	64	0.499
9MSI.A	66	β	63	0.499

[a] SCOP classification.[22]
[b] A soluble domain from a larger membrane protein.

For a maximal chain length $n = 128$ divided into sequential blocks of $b = 8$ residues, there are $128/8 = 16$ blocks, resulting in $15 \times 16/2 = 120$ pairs of blocks, corresponding to 120 terms in Eq. (2) and the same number of dimensions of the conformation space. The initial, all-encompassing range consists of lower and upper bounds on the 120 dimensions derived from geometric considerations

and physical limitations on the maximal packing density of polypeptide chains, the maximal extent of polypeptide chains, etc.[12] Clearly, shorter chains would involve a lower dimensionality of the initial range, but it is computationally convenient to treat all the training set proteins with one range. This is done by the equivalent of "zero filling" in Fourier analysis and signal processing, where extra "virtual" residues are added to the C-terminus of each protein to bring its chain length up to 128. These extra residues are placed in an extended chain attached to the C-terminal "real" residue and pointing radially out from the protein's center of mass. They are given an imaginary 21st amino acid type that is taken to be noninteracting, so they do not contribute to the block pair contents $\boldsymbol{t}_{IJ}$ in Eq. (3). Thus the sum of the components of $\boldsymbol{t}_{IJ}$ is the product of the number of real residues in blocks I and J. While the extra residues have no effect on the energetic side of the model, they do increase the conformational entropy of the smaller proteins in Table I, such as 1ENH, which has 54 residues. Better treatment of variable chain length would be a topic for future improvements.

Having only a single range corresponds to a trivial model where there is no discrimination between folded and unfolded states of the protein. At the opposite extreme, consider a "shrink-wrap" model consisting of 32 very small, nonoverlapping ranges, each centered around one of the training set PDB structures. Now if the κ cutoff in Eq. (3) is chosen to be very large, all pairs of blocks in all 32 ranges are always in contact, and each protein would have the same energy in each range. At the opposite extreme, a very small value for κ would give the same zero energy value for all proteins in all ranges. It turns out there is a rather narrow range for κ in between such that for each of the training proteins, the energy of the native range corresponds to a distinct polynomial in the 210 residue type pairs compared to those of the 31 respective nonnative ranges. We find that for $\kappa = 26096.7$ Å^2 it is possible to adjust the 210 components of $\boldsymbol{a}$ so that the energy of the native range of each of the 32 training proteins is lower (more favorable) than the energy of the 31 other nonnative ranges. This value of κ is used in all that follows.

Of course the shrink-wrap model is not satisfactory for protein folding because it neglects all conformations that are not tightly folded. To get a set of ranges that still cover all possible conformations while including less disallowed volume than the initial range, our general procedure is illustrated in Figure 3. Starting with the single initial range (the large heavy rectangle in the figure), choose a dimension and a value along that dimension between the upper and lower bounds at which to split the range into two subranges. The two subranges may often be contracted in other dimensions while still containing all their allowed conformations. In this work we have used no more than triangle inequality level reasoning to perform the contractions, which is more involved in cluster distance geometry than in standard distance geometry.[12] Thus one subrange has its upper bound lowered to the split value in the chosen dimension, and this implies that some other pairs of blocks nearby in sequence now also have reduced upper bounds to a lesser extent. Similarly,

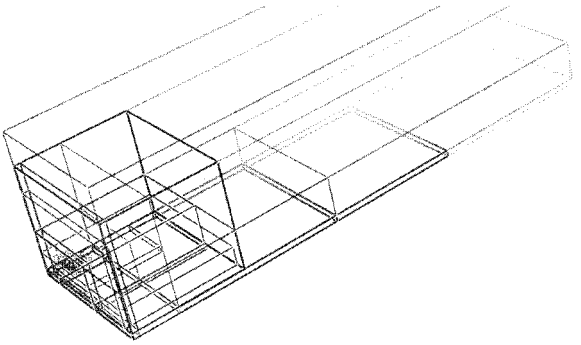

FIGURE 4 Projection onto three dimensions of the 71 regions in 120 dimensions. Ranges appear to overlap, but this is an artifact of the projection. The largest box contains all the others and represents all chain conformations, while small boxes near the lower left corner correspond to native ranges.

the other subrange has a raised lower bound in the chosen dimension, and this implies some increases in the lower bounds of other dimensions. The process continues until some stopping criterion has been reached, such as no range still containing at least one of the training set protein native structures has a dimension available for splitting that is adequately long. Such a criterion has the advantage of not further subdividing ranges corresponding to conformations far from any of the native structures. In any case, note that the union of the final set of ranges should still cover all geometrically allowed conformations, but it always includes some amount of disallowed volume, as illustrated in Figure 3.

Specifically, the range-splitting algorithm we have found to be most satisfactory is as follows. (1) Start with the initial range that encloses all conformations for the given chain length. (2) For each range that still encloses at least one of the training protein structures, consider each dimension in turn that corresponds to distances between real (not virtual zero fill) residues of those proteins. If $\kappa > D_{IJ}$ for all the natives and $U_{IJ} > 5\kappa$, then split that dimension at κ and go to step 4. (3) Alternatively, if the range encloses more than one native structure and that dimension has native D_{IJ} values both above and below κ, then split that dimension at κ and go to step 4. (4) If a range was split according to steps 2 or 3, contract the upper and lower bounds of each of the new ranges using triangle inequality bound smoothing. Then reexamine the list of ranges again in step 2. Otherwise, there was no opportunity to split any range, and the procedure is finished.

Note that the restriction in step 2 to dimensions involving real residues has two beneficial effects. First, the dimensions involving virtual residues are distance intervals that are left intentionally broad, rather than pointlessly specifying the conformation of virtual residues in more detail. Secondly, any splitting in step 3 will differentiate between conformations that will potentially have different interaction energies, so that subsequent adjustment of energy parameters will enable a given protein to prefer one range over the other. Another way to view this

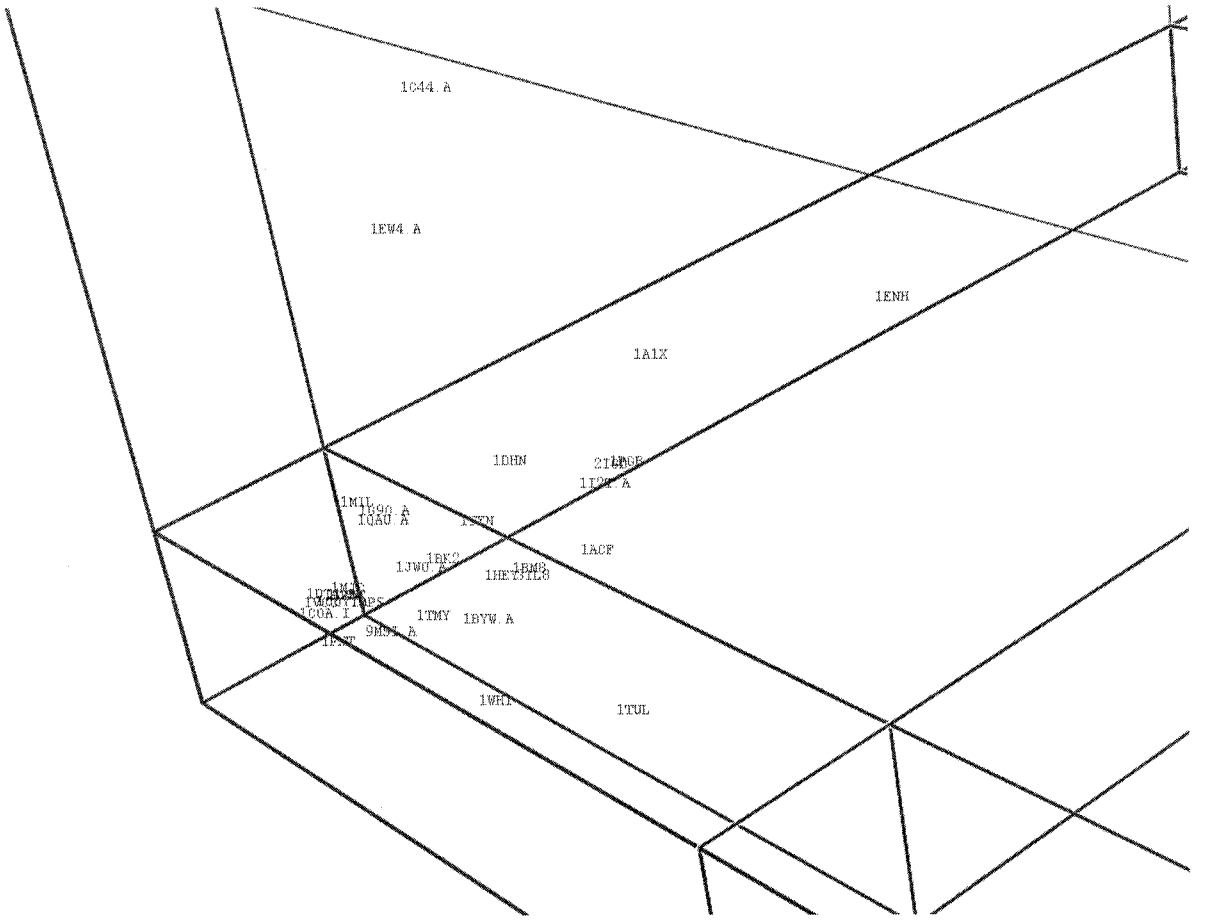

FIGURE 5 Close-up of the native corner of Figure 4, showing labels representing the native conformations of the training set proteins.

effect is that when the final set of ranges are projected onto the subspace of dimensions corresponding to real residues of a protein in the training set, the native structure always falls into exactly one range, rather than the projected ranges overlapping each other.

Applying this algorithm to the training set in Table I using $\kappa = 26096.7$ Å^2 produces a set of 71 ranges where the first 51 correspond to large volumes of less compact conformations, and the last 20 enclose the 32 training proteins. While it is hard to draw satisfactory pictures of 120 dimensional rectangles, Figures 4 and 5 illustrate the general idea. Table I tells which range each native structure falls inside. Note that some such ranges include only one native, while range 71 includes eight of them. Extensive testing has never discovered a real polypeptide conformation that is not contained in one of the ranges or falls within two of them, so these 71 ranges are apparently nonoverlapping, cover all allowed conformations, and err only by including some regions of conformation space that correspond to no three-dimension polypeptide conformation.

Clearly the final set of 71 ranges depends on the training set. In part, they are used to describe the full conformation space by devoting greater detail to the vicinity of the structures of real, globular proteins. Three-dimensional chain conformations also have substantial nonlinear constraints on the D_{IJ} values beyond the triangle inequality relations used in the range-splitting procedure. Thus the training set proteins also serve to focus attention on geometrically realizable structures. The native conformations of all other proteins we have examined always fall into one and only one of the 71 ranges, but sometimes they lie in one of the larger ranges not associated with any of the training set proteins. In future work, we would like to determine sets of ranges without reference to a particular training set in order to make the description of conformation space more general.

Partition Function

Neglecting the momentum terms in the full Hamiltonian, the classical canonical partition function consists of an integral of the Boltzmann factor integrated over all conformational degrees of freedom. In our model, this conveniently breaks up into a sum over each of the 71 ranges, and the integral over the 120 dimensions becomes a product of 120 integrals

Table II Energy Parameters

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	−0.00																			
S	−0.62	0.20																		
T	−0.67	0.31	−0.04																	
P	0.12	−0.69	0.46	−0.27																
A	−0.27	−0.93	0.47	0.26	−0.41															
G	−0.14	−0.68	−0.10	−1.61	0.44	−0.08														
N	−0.00	0.32	0.74	−0.57	−0.14	−1.45	1.25													
D	−0.47	−0.48	−0.76	0.04	−0.01	−0.24	−0.18	0.92												
E	0.26	−0.84	0.15	0.39	−0.20	0.82	−0.87	1.12	−0.16											
Q	−0.58	−0.15	0.01	−1.13	−0.04	0.42	−0.84	−0.24	−0.42	0.65										
H	−0.18	−0.92	0.50	−0.73	0.55	−0.73	−0.10	1.27	0.36	0.18	0.26									
R	−0.11	0.19	−0.34	−0.72	0.46	−0.35	−0.68	0.17	−0.07	−0.21	0.11	−0.26								
K	0.91	−0.37	−0.10	−0.15	0.14	0.31	−0.17	−0.69	−0.53	−0.11	−0.39	0.23	−0.06							
M	−0.20	−0.52	−0.19	−0.42	−0.77	0.97	1.13	−0.49	0.09	0.33	0.36	−1.62	0.43	−0.06						
I	−0.50	0.02	0.18	0.19	0.59	−0.61	1.25	−0.03	−0.86	0.40	0.34	−0.18	0.46	−0.08	−0.56					
L	−0.27	0.72	0.24	0.55	−0.90	0.28	−0.93	−0.22	−0.11	0.21	−0.50	0.27	−0.53	0.63	−0.25	0.28				
V	0.33	0.23	−0.33	0.32	−0.52	0.16	0.26	0.07	0.03	−0.45	−0.07	0.21	−0.12	−0.27	−0.79	−0.43	0.68			
F	0.07	0.49	−0.48	−0.99	−0.26	−0.04	−0.22	0.10	0.82	0.33	−0.46	0.25	0.05	0.93	0.27	−0.76	−0.18	0.79		
Y	−0.10	−0.39	0.25	1.57	−0.30	0.64	−0.16	−0.37	−1.01	−0.49	0.40	0.93	0.39	−0.63	−0.66	−0.07	−0.95	−0.61	0.43	
W	−0.10	−0.64	−0.05	0.14	0.88	1.04	−0.59	0.12	0.12	−0.87	0.52	0.02	−0.18	0.21	−0.35	−0.26	−0.18	−0.83	0.49	−0.30

Table III Correctly Predicted Proteins

PDB Entry	Type[a]	Native Range	Z_{nat}/Z_{tot}	% Sequence	Identity to[b]
1A8O	α	7	1	21	1I2T.A
1AA3	$\alpha + \beta$	71	1	21	1UBI
1ACP	α	7	1	21	1DT4.A
1APF	Small protein	37	1	21	1BK2
1AVS.B	α	7	1	23	1PHT
1AY7.B	α/β	71	0.999	21	1PTF
1AYJ	Small protein	71	1	21	1I2T.A
1B00.A	α/β	71	1	27	1TMY
2B3I.A	β	71	1	22	1TEN
1B4B.A	$\alpha + \beta$	71	1	18	1COA.I
1BB8	$\alpha + \beta$	7	1	18	1I2T.A
3BBG	Small protein	25	1	15	1JWO.A
1BBL	α	25	1	23	2IGD
1BBO	Small protein	7	0.982	18	3IL8
2BBV.D	α	2	0.998	7	1ENH
1BDS	Small protein	37	1	16	1BK2
1BFG	β	71	1	21	1WHI
1BHU	β	71	1	20	1HEY
1BIK	Small protein	71	1	17	1BYW.A
2BJX.A	α/β	71	0.611	21	1DHN
1BK8	Small protein	71	1	18	3IL8
1BLU	$\alpha + \beta$	71	1	23	1DT4.A
1BQX.A	$\alpha + \beta$	71	1	21	1BYW.A
1BUH.B	$\alpha + \beta$	7	1	24	1BK2
1BUS	Small protein	71	1	28	3IL8
1BYF.A	$\alpha + \beta$	71	1	22	1HEY
1C01.A	β	71	1	20	1MJC
1C55.A	Small protein	25	1	16	1OPS
1C75.A	α	71	1	22	1BM8
1CCF	Small protein	7	1	19	1PHT
1CDQ	Small protein	71	1	23	3IL8
1CL4.A	Small protein	16	1	16	1OPS
1CLF	$\alpha + \beta$	71	0.954	23	1PGB
1COU.A	Small protein	7	1	21	1BYW.A
1CPZ.A	$\alpha + \beta$	71	0.959	21	1COA.I
1CXW.A	Small protein	71	0.972	21	1JWO.A
1D6R.I	β	71	1	21	3IL8
1DAV.A	α	7	0.999	24	1COA.I
1DD3.A	α	7	1	20	1WHI
1DD4.D	α	7	1	21	1PGB
1DPT.A	$\alpha + \beta$	71	1	23	1MIL
1DUR.A	$\alpha + \beta$	71	1	20	1COA.I
1DX5.I	Small protein	7	1	18	1BYW.A
1DZ3.A	α/β	71	1	28	1TMY
1E53.A	Small protein	71	0.999	18	1UBI
1E8Q.A	Small protein	37	1	15	2IGD
1E8R.A	Small protein	71	1	20	9MSI.A
1E9T.A	Small protein	17	1	19	1BK2
1EGP.A	$\alpha + \beta$	16	0.762	22	1COA.I
1EJA.B	Small protein	7	1	20	1OPS
1EP7.A	α/β	71	0.977	24	1TMY
1ERD	α	25	0.868	16	1BK2
2ERL	α	25	1	17	1UBI
1ERV	α/β	71	1	20	1C44.A

Table III (*Continued from the previous page*)

PDB Entry	Type[a]	Native Range	Z_{nat}/Z_{tot}	% Sequence	Identity to[b]
1F4I.A	α	25	1	19	1OPS
1F5Y.A	Small protein	7	1	19	1DT4.A
1F66.G	α	7	1	20	1ACF
1FBQ.B	α	71	0.999	18	1PTF
1FE4.A	$\alpha + \beta$	71	1	19	1VCC
1FI6.A	α	7	1	21	1MJC
1FM0.D	$\alpha + \beta$	71	1	23	1BYW.A
1FMY.A	Small protein	25	1	15	1ENH
2FN2	Small protein	71	1	20	9MSI.A
1FO5.A	α/β	63	0.948	23	1PTF
1FRE	Small protein	25	1	18	1BK2
1FSB	Small protein	25	1	16	1BK2
1FVS.A	$\alpha + \beta$	71	0.765	18	1PTF
1FXR.A	$\alpha + \beta$	71	0.982	18	1BYW.A
1G4D.A	α	71	1	19	1OPS
1G6M.A	Small protein	71	1	24	1ENH
8GCH	β	1	0.62	8	9MSI.A
1GD0.A	$\alpha + \beta$	71	1	18	1DHN
1GMM.A	β	71	0.95	21	1TUL
1GPS	Small protein	37	1	17	1PHT
1GPT	Small protein	37	1	18	3IL8
1GYZ.A	α	71	1	18	3IL8
1H75.A	α/β	71	1	20	1TMY
1H9E.A	α	71	1	25	1COA.I
1HAF	Small protein	25	0.747	24	3IL8
1HCD	β	71	1	20	1C44.A
1HDF.A	β	71	1	24	1G9O.A
1HEV	Small protein	25	1	16	3IL8
1HH5.A	α	7	1	21	1DT4.A
1HPI	Small protein	7	1	21	1I2T.A
1HY9.A	Small protein	25	1	17	9MSI.A
1HYK.A	Small protein	37	1	14	1MIL
1I0V.A	α/β	7	0.995	22	1TEN
1I2U.A	Small protein	37	1	17	1MJC
1ICA	Small protein	25	1	16	1G9O.A
1IFY.A	α	25	0.999	18	3IL8
1IGL	Small protein	7	0.928	23	1JWO.A
1IIB.A	α/β	71	1	21	1C44.A
1IQT.A	$\alpha + \beta$	71	1	21	1BK2
1IYV	β	52	1	24	1PTF
1J0T.A	α	7	1	18	1BYW.A
1J6Q.A	β	16	0.515	23	1G9O.A
1J75.A	α	71	1	20	1G9O.A
1J7M.A	Small protein	71	1	21	1JWO.A
1JBG.A	α	71	1	21	1EW4.A
1JE9.A	Small protein	71	1	23	1BK2
1JK4.A	β	71	1	24	1MJC
1JKZ.A	Small protein	37	0.998	23	2IGD
1JRH.I	β	11	1	20	3IL8
1KJK.A	$\alpha + \beta$	37	1	25	1DT4.A
1KN6.A	$\alpha + \beta$	71	1	19	1COA.I
1KV4.A	peptide	2	0.5	21	2IGD
1KVZ.A	$\alpha + \beta$	71	1	21	1DHN
1KW4.A	α	7	1	24	1MJC

Table III (*Continued from the previous page*)

PDB Entry	Type[a]	Native Range	Z_{nat}/Z_{tot}	% Sequence	Identity to[b]
1KX1.A	Small protein	71	1	20	1UBI
1L3Y.A	Small protein	25	0.912	18	3IL8
1L4V.A	Small protein	25	1	18	1BK2
1L5D.A	β	7	0.997	20	1WHI
1LD6.A	Small protein	71	1	17	1OPS
1LDL	Small protein	7	0.999	16	1MJC
1LNG.A	$\alpha + \beta$	71	0.984	23	1UBI
1LSW.A	$\alpha + \beta$	71	0.998	20	1HEY
1M5T.A	α/β	71	1	26	1TMY
1M7T.A	α/β	71	1	20	1C44.A
1M8A.A	$\alpha + \beta$	71	1	25	3IL8
1MFI.A	$\alpha + \beta$	71	1	23	1ACF
1MHD.A	$\alpha + \beta$	71	1	20	1JWO.A
2MHR	α	17	1	21	1TUL
2MOB.A	$\alpha + \beta$	71	1	23	1UBI
1MVO.A	α/β	71	1	29	1TMY
1MYN	Small protein	37	0.993	18	1PGB
1NXB	Small protein	71	0.633	18	2IGD
3OVO	$\alpha + \beta$	2	1	9	3IL8
1PDN.C	α	7	0.975	23	1G9O.A
1POU	α	9	0.999	23	3IL8
2PSP.A	Small protein	71	1	17	1BYW.A
1QJG.A	$\alpha + \beta$	8	1	20	1DHN
1QR5.A	$\alpha + \beta$	71	1	66	1PTF
1QUQ.D	β	7	1	21	1HEY
1R69	α	71	1	21	9MSI.A
1REG.X	$\alpha + \beta$	7	0.999	22	1JWO.A
1RMD	Small protein	71	1	20	1EW4.A
1RNV	Small protein	2	0.998	8	2IGD
2RSL.B	α/β	71	1	20	1EW4.A
1SNB	Small protein	71	1	18	1DT4.A
1T1D.A	$\alpha + \beta$	71	1	21	1G9O.A
1TFI	Small protein	7	1	19	1VCC
2TGI	Small protein	7	1	20	1PHT
1TGX.A	Small protein	7	1	24	3IL8
1TNT	α	71	1	19	1G9O.A
1TPN	Small protein	7	1	17	1A1X
2TRX.A	α/β	71	1	21	1BYW.A
1WKT	β	71	0.999	20	1JWO.A

[a] SCOP classification.[22]

[b] Percent sequence identity of most similar protein from the training set.

over one dimension each, because the potential energy in Eq. (2) was built to be a sum of terms, each depending on one of the conformational variables. Thus

$$Z_{tot} = \sum_{r=1}^{71} \prod_{I<J} \int_{L_{rIJ}}^{U_{rIJ}} \exp[-\beta E(D_{IJ}, \boldsymbol{t}_{IJ})] dD_{IJ} \qquad (4)$$

where $\beta = (k_B T)^{-1}$ for some arbitrary temperature T. The partition function depends on the protein's amino acid sequence via the interblock residue type pair composition vectors $\boldsymbol{t}_{IJ}$. Because E is a simple square well, the integral for a particular range r and dimension IJ is

$$\int_{L_{rIJ}}^{U_{rIJ}} \exp[-\beta E(D_{IJ}, \boldsymbol{t}_{IJ})] dD_{IJ} = \begin{cases} U_{rIJ} - L_{rIJ} & \text{for } \kappa < L_{rIJ} \\ (U_{rIJ} - L_{rIJ}) \exp[-\beta(\boldsymbol{a} \cdot \boldsymbol{t}_{IJ})] & \text{for } \kappa > U_{rIJ} \\ (U_{rIJ} - \kappa) + (\kappa - L_{rIJ}) & \\ \quad \times \exp[-\beta(\boldsymbol{a} \cdot \boldsymbol{t}_{IJ})] & \text{otherwise} \end{cases} \qquad (5)$$

depending on the relative values of κ and the lower and upper bounds on D_{IJ}.

We take the partition function of the macroscopic native state to be just the one term in Eq. (4) where $r = r_{nat}$, the one range containing the PDB structure of the protein. The partition function of the denatured state is then the sum over the other 70 ranges. The Helmholtz free energy of any macroscopic state is $A = -\beta^{-1} \ln Z$. At the midpoint of thermal denaturation, $A_{nat} = A_{den}$, which is equivalent to $Z_{nat} = Z_{den}$ or $Z_{nat}/Z_{tot} = 0.5$, since the native range is included in Z_{tot}.

Optimization of Parameters

The 210 adjustable energy parameters in $\boldsymbol{a}$ were determined by requiring that the native free energy of all 32 training proteins be at least equal to the denatured free energy, and otherwise the sum of the squares of the parameters should be minimal. The straightforward constraints are that $Z_{p,nat}/Z_{p,tot} > 0.5$ for each protein p in the training set. However, there are technical problems with floating point overflow and underflow when multiplying 210 factors together to evaluate the partition function associated with one range, so the calculations were actually done in terms of logarithms of the partition functions. This was greatly facilitated by a function in MOE's SVL programming language[13] called logaddexp for summing numbers of disparate magnitudes. Denoting this function by S, it calculates $S(i = 1, \ldots, n, x_i) = \ln \Sigma \exp(x_i)$ in terms of logarithms. For example, from Eq. (4) we have $\ln Z_{tot} = S(r = 1, \ldots, 71, \Sigma_{I<J} \ln \int \exp(-\beta E) dD_{IJ})$. Then the constraints were satisfied by local minimization of penalty terms

$$F_{obj} = \sum_p \begin{cases} (\ln Z_{p,nat} - \ln Z_{p,tot} - \ln 0.5)^2 \\ \quad \text{if } Z_{p,nat}/Z_{p,tot} < 0.5 \\ 0 \qquad \text{otherwise} \end{cases} \tag{6}$$

summing over all 32 polypeptide chains p in the training set. F_{obj} can be reliably optimized even starting from initial random values of the components of $\boldsymbol{a}$ in the range zero to one, using a truncated Newton unconstrained local optimization procedure with analytical gradiant. We were not troubled by convergence to local minima where F_{obj} was significantly greater than zero.

Table I shows that even short of final convergence of the minimization, the constraints are essentially satisfied for all proteins, and they are even slack for 1UBI and 1VCC. For five of them (1BK2, 1BYW.A, 1JWO.A, 1PTF, and 1TMY) the total partition function is dominated by the native range and one nonnative range that has a slightly greater partition function. Otherwise, the native range predominates to any desired degree at lower temperatures.

The resulting interaction energy parameters are shown in Table II. While many of them make some intuitive sense, such as negative (favorable) values for hydrophobic–hydrophobic interactions in the lower right corner of the table, there are many exceptions. Presumably a larger training set and/or a more detailed energetic model would produce parameters having seriously interpretable values. Yet the fact that most training set proteins have barely stable native states suggests that these are the smallest parameter magnitudes that can account for their stabilities. The real test is whether the final model of ranges, energy function, and adjustable parameters has any predictive power.

Prediction of Short Polypeptide Chains

The MOE[13] molecular modelling software comes with a nonredundant subset of the polypeptide chains in PDB chosen to be those relatively precisely determined structures having little missing coordinate data, and generally less than 90% sequence identity between any pair of chains. Out of this list of 4986 PDB entries, we selected those consisting of only a single polypeptide chain having no more than 128 residues, all of them having the 20 standard types, and all of them having coordinates for their C^{α} atoms. Each of these 1540 chains fell inside one of the 71 conformational ranges, and for 174 of these, the native range's free energy was lower than that of all 70 nonnative ranges' free energies ($Z_{nat} > Z_{non}$). If all ranges were equally likely to be a protein's native range, one would expect only 1540/71 = 22 correct predictions. These results are also not simply due to sequence homology, because of the 174 correct predictions, 146 of them have less than 30% sequence identity (after optimal gapped sequence alignment) to any of the training proteins. These 146 are listed in Table III. For almost all of these predictions, the native state is strongly favored over the denatured in free energy ($Z_{nat}/Z_{tot} = 1$), and range 71 is the most frequently occurring native range, although a variety of different SCOP fold types are represented in it.

When a given protein is predicted to have its native structure in the correct range, it is something of an ab initio protein fold prediction on the basis of the given amino acid sequence and the statistical mechanical model that encompasses all possible conformations of a single chain having no more than 128 residues. Having divided all conformations into only 71 ranges, however, suggests that even the smaller ranges containing the training set proteins are still rather broadly defined. Consider one of the successful predictions, PDB entry 1AA3, an NMR structure of a single chain of 63 residues without crosslinks or substantial ligands. Deleting the last seven residues, it can be viewed as seven blocks of eight residues each, and just as in Table III with zero filling, its conformation falls within range 71. Taking the limits on the corresponding dimensions of that range as distance constraints, we derive $7 \times 6/2 = 21$ upper and lower bound constraints on the interblock distances, D_{IJ}. Using cluster distance geometry metric matrix embedding,[12] we found that these 42 constraints are easy to satisfy, and the sampling of five structures produced had root mean square deviations (RMSDs) to the native ranging from 8.3 to 11.5 Å. They were all fairly compact, but they had large root mean square deviations in coordinates of the centers of mass of the seven blocks after optimal superposition onto the native structure's corresponding centers of mass. This is all consistent with the ranges being only very general descriptions of sets of conformations.

Agreement with the PDB structure is gratifying, but actually most polypeptide chains in PDB entries have features that may greatly affect their conformation but are outside the scope of the current model, namely multiple polypeptide chains, associated polynucleotide chains, disulfide crosslinks, and a wide variety of substantial ligands or cofactors. Therefore it is not surprising that many chains, taken in isolation, have conformations that do not agree with these calculations.

DISCUSSION AND CONCLUSIONS

We have described a version of SMEUSE based on cluster distance geometry. Even representing polypeptide chains as blocks of eight sequentially adjacent residues, it is possible to construct a very simple energy function involving interactions between blocks such that a set of 32 proteins have native state free energies more favorable than that of the denatured state. Then 146 other nonhomologous proteins are correctly predicted to favor their native states. There are several possible explanations for not every protein in PDB being correctly predicted. One is that most PDB entries involve significant contributions to their stabilities that are not included in this model, such as disulfide crosslinks, quaternary structure, and binding of large ligands. Another explanation is that the model is oversimplified and/or views protein folds at too low a resolution.

Many of the limitations of the current model can be readily improved while still keeping its computational complexity under control. Chains longer than 128 residues can be accommodated with increased dimensionality of the ranges. Multiple polypeptide chains can be treated by a different set of ranges that have no chain connectivity constraints between chains. Different disulfide bridge arrangements can be included by constructing special ranges that include the corresponding low upper distance bounds. The set of all possible conformations can be described more accurately by larger numbers of more narrowly bounded ranges. Another way is to reduce the number of residues in a block. More sophisticated energy terms can be incorporated in order to better reflect the exact amino acid sequence and the environment of the residues. In any case, the current model or such future improvements maintain the key features of SMEUSE: all conformations are included, and the partition function can be readily evaluated so that the thermodynamics of protein folding can be confidently modelled.

REFERENCES

1. Guo, Z.; Brooks, C. L.; Boczko, E. M. Proc Natl Acad Sci USA 1997, 19, 10161–10166.
2. Vorobjev, Y. N.; Hermans, J. Protein Sci 2001, 10, 2498–2506.
3. Zagrovic, B.; Sorin, E. J.; Pande, V. J Mol Biol 2001, 313, 151–169.
4. Borreguero, J. M.; Dokholyan, N. V.; Buldyrev, S. V.; Shakhnovich, E. I.; Stanley, H, E. J Mol Biol 2002, 318, 863–876.
5. Skolnick, J.; Kolinski, A. J Mol Biol 1990, 212, 787–817.
6. Hansmann, U. H. E.; Okamoto, Y. Ann Rev Comp Phys VI 1999, 129–157.
7. Zhang, H. Proteins 1999, 34, 464–471.
8. Mitsutake, A.; Sugita, Y.; Okamoto, Y. Biopolymers 2001, 60, 96–123.
9. Zhang, Y.; Kihara, D.; Skolnick, J. Proteins 2002, 48, 192–201.
10. Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. Curr Opin Struct Biol 2003, 13, 168–174.
11. Wang, J.; Crippen, G. M. Biopolymers 2004, 74, 214–220.
12. Crippen, G. M. J Comput Chem 2004, 25, 1305–1312.
13. Molecular Operating Environment (MOE), Chemical Computing Group, Inc., http://www.chemcomp.com.
14. Crippen, G. M. J Comput Chem 1992, 13, 351–361.
15. Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation; Wiley: New York, 1988.
16. Ghose, A. K.; Crippen, G. M. J Comput Chem 1985, 6, 350–359.
17. Wildman, S. A.; Crippen, G. M. J Mol Graph Mod 2002, 21, 161–170.
18. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235–242.
19. Hobohm, U.; Sander, C. Protein Sci 1994, 3, 522–524.
20. Maiorov, V. N.; Crippen, G. M. J Mol Biol 1992, 227, 876–888.
21. Kabsch, W. Acta Cryst 1978, A34, 827–828.
22. Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. Nucleic Acid Res 2002, 30, 264–267.

Reviewing Editor: Dr. David A. Case