Protein Science (1996), 5:254-261. Cambridge University Press. Printed in the USA.

Folding proteins with a simple energy function and extensive conformational searching

KAIZHI YUE AND KEN A. DILL

Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143

(RECEIVED October 17, 1995; ACCEPTED December 7, 1995)

Abstract

We describe a computer algorithm for predicting the three-dimensional structures of proteins using only their amino acid sequences. The method differs from others in two ways: (1) it uses very few energy parameters, representing hydrophobic and polar interactions, and (2) it uses a new "constraint-based exhaustive" searching method, which appears to be among the fastest and most complete search methods yet available for realistic protein models. It finds a relatively small number of low-energy conformations, among which are native-like conformations, for crambin (1CRN), avian pancreatic polypeptide (1PPT), melittin (2MLT), and apamin. Thus, the lowest-energy states of very simple energy functions may predict the native structures of globular proteins.

Keywords: conformational search; energy potential; protein folding; structure prediction

During the past 20 years, there have been several important advances in computer algorithms intended to predict native three-dimensional structures of globular proteins from their amino acid sequences (Levitt & Warshel, 1975; Kuntz et al., 1976; Wilson & Doniach, 1989; Skolnick & Kolinski, 1990; Covell, 1992; Sippl et al., 1992; Vajda et al., 1993; Covell, 1994; Hinds & Levitt, 1994; Kolinski & Skolnick, 1994; Monge et al., 1994; Wallqvist & Ullner, 1994; Boczko & Brooks, 1995). These methods assume that a native protein structure is a balance of many different interactions, usually characterized using hundreds to several thousands of "knowledge-based" energy parameters derived from databases of known protein structures. Based on the thermodynamic hypothesis (Anfinsen, 1973) that the native three-dimensional structure of a protein is the state of lowest free energy, these algorithms explore many different protein conformations by sampling methods, such as Monte Carlo, molecular mechanics, or molecular dynamics, to find the most stable conformations. Each such method correctly predicts a few protein structures but misses many others.

A much simpler approach has recently emerged, based on using far fewer energy parameters that represent simple physical quantities and are not derived from knowledge bases of protein structures (see Srinivasan & Rose, 1995; Sun et al., 1995). This approach is attractive because it should be much easier to learn how to improve few-parameter models than highly parameterized ones, and the number of successful or partially successful predictions from these simple models is arguably comparable to those from the more complex models. However, it is not clear whether either the method of Srinivasan and Rose (1995) or that of Sun et al. (1995) is consistent with the thermodynamic hypothesis, because Sun et al. use artificial native structure restraints and Srinivasan and Rose use a hierarchical method to overcome computer conformational searching limitations. Here we develop an algorithm called Geocore that uses an equally simple energy function combined with a new and powerful constraint-based exhaustive searching method. We show that: (1) native-like protein structures can be found as the thermodynamically stable states of such simple physical energy functions, and (2) exhaustive methods, normally dismissed as being computationally impractical for proteins, appear to be viable alternatives to sparse sampling methods such as Monte Carlo and molecular dynamics.

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

Description of Geocore

As all computer folding methods, Geocore has three aspects.

Chain representation

Amino acids are represented at the united-atom level, except polar hydrogens, which are explicit. Each atom or united atom is a hard sphere with its appropriate van der Waals (vdW) radius, R_{vdW}. Because van der Waals radii are larger than those implied by the minimum contact distances observed in proteins (Cantor & Schimmel, 1980), we soften the potential using a steric allowance, $R_a(A, B)$ for atoms A and B, chosen partly by how much we wish to restrict the computer search; it is typically 0.2–0.3 Å per instance, and summed together adds up to

a maximum of 5 Å total allowance for the whole protein. Thus, the minimum contact distance between atoms A and B is $R_{vdW}(A) + R_{vdW}(B) - R_a(A, B)$. Backbone conformations are represented using a discrete set of dihedral angles (ϕ, ψ), as in the rotational isomeric state (RIS) model of polymers (Flory, 1969). The peptide bond is assumed to be strictly planar. The ϕ, ψ angle preferences of the different amino acids are taken from our survey of 25 small proteins in the Protein Data Bank (PDB)—1CRN, 1ECA, 1MBD, 1NXB, 1REI, 1RNS, 1SN3, 2ACT, 2CCY, 2CYP, 2LZM, 2MHR, 2MLT, 2PTN, 2SNS, 2WRP, 3B5C, 3DFR, 3INS, 4CPV, 4FXN, 4GCR, 5CYT, 6PTI, 7RSA—and are given in Table 1. Table 1 shows that the maximum number of ϕ, ψ choices depends on the amino acid: for example, glycine has six, proline has three, others have four or five.

We search all of these rotational isomers uniformly, provided they are consistent with steric constraints (see below), and thus we give no preference to any particular isomer. The use of discrete rotamers trades off the flexibility observed in real proteins with the computational need to limit the conformational searching. A premise of Geocore is that the ϕ, ψ angles observed in proteins are determined more by nonlocal hydrophobic and hydrogen bonding interactions and steric restraints than by ϕ, ψ angle energetic preferences. Because there are no dihedral angle preferences in Geocore, local interactions can be regarded as providing a set of options for the conformational search rather than giving direction to the search.

Potential function: (a) Hydrophobic interactions

Geocore assumes that hydrophobic interactions are an important sequence-specific structure-causing force (reviewed in Dill, 1990; Dill et al., 1995). It has been found in lattice models that maximizing the number of contacts between nonpolar groups is a very strong pruning constraint in the conformational search to find protein-like native states (Yue & Dill, 1995; Yue et al., 1995). Here we generate conformations with maximal and near-maximal pairwise shared nonpolar surface among nonpolar atoms, which is a good approximation to global minimization of exposed nonpolar surface area.

We define polar and nonpolar united atoms by their heavy atoms: carbon and sulfur are nonpolar; nitrogen and oxygen are polar. Native structures of proteins have minimal solvent-exposed nonpolar surface areas (Eisenberg et al., 1984). As lowest-energy states, our algorithm seeks conformations with minimal exposed nonpolar surface area by maximizing the pairwise shared nonpolar surface area. A nonpolar atom is defined as exposed if it contacts solvent or polar groups (Yue & Dill, 1995). When two nonpolar atoms i and j share a surface area σ, the interaction energy is $E(i, j) = -c\sigma$, where c is a positive constant such that at the closest separation, $E(i, j) = -1$. This defines an HH contact. Following Lee and Richards (1971), $E(i, j) = 0$ when the distance between the atoms is greater than or equal to the diameter of a water molecule. In the spirit of De la Cruz et al. (1992), we define the shared surface area as the solid angle from which sphere A "sees" sphere B. As shown in Figure 1, the shared surface area is defined by the cone formed by lines from the center of sphere A to become tangent to sphere B. If the solid angle of the cone is Ω and the radius of atom A is r, then the shared surface is $4\pi r^2 \times (\Omega/4\pi) = \Omega r^2$, where $\Omega = 2\pi[1 - \sqrt{1 - (r'/d)^2}]$, d is the distance between the atom centers, and r' is the radius of atom B. Based roughly on oil/water partition experiments, we set the energy per HH contact to be -0.7 kcal/mol. When two nonpolar amino acids are adjacent, they can have multiple HH contacts.

Table 1. *The ϕ, ψ angle options used by Geocore*[a]

	ϕ, ψ	Fr	ϕ, ψ	Fr	ϕ, ψ	Fr	ϕ, ψ	Fr	ϕ, ψ	Fr	ϕ, ψ	Fr
C	247, 144	63	287, 324	37	63, 21	1	189, 90	1				
M	296, 322	31	244, 140	23	63, 32	1	211, 62	1				
F	297, 317	71	264, 132	28	217, 158	12	251, 348	9	67, 31	4	131, 164	2
I	292, 319	83	247, 132	75	259, 33	3	2, 127	1	80, 283	1	216, 186	1
L	293, 323	147	264, 136	84	123, 303	2	46, 85	1	64, 18	1	217, 85	1
V	296, 318	88	250, 132	83	243, 342	8	198, 164	5	68, 1	1		
W	296, 322	25	248, 149	22	249, 22	7						
Y	249, 138	74	289, 325	45	241, 65	8	72, 27	5	39, 120	2	186, 332	1
A	295, 324	199	283, 143	42	209, 154	23	248, 25	18	61, 310	1		
G	292, 325	62	75, 198	35	152, 179	34	271, 173	33	96, 349	29	270, 25	9
T	254, 142	95	286, 325	84	207, 188	2	232, 348	2	168, 55	1	228, 90	1
S	246, 144	100	287, 331	89	47, 55	8	48, 287	1	157, 347	1	244, 236	1
E	293, 322	145	263, 140	48	58, 43	4	242, 193	2	314, 64	1	316, 263	1
N	293, 322	55	246, 134	55	66, 31	15	43, 229	2	288, 225	1		
Q	292, 327	80	251, 143	48	64, 247	2	244, 350	2	262, 226	2	8, 87	1
D	290, 327	90	273, 134	45	205, 176	7	210, 87	7	62, 34	6	19, 280	1
H	291, 324	33	256, 129	28	239, 35	7	234, 346	3	52, 44	2	260, 247	1
R	294, 322	71	250, 144	39	232, 72	2	61, 39	1	148, 186	1		
K	294, 324	146	257, 137	84	69, 58	9	246, 18	5	275, 69	4	123, 242	2
P	294, 150	69	297, 330	60	357, 95	1						

[a] The "Fr" columns give the frequencies of the ϕ, ψ angles observed in the set of test proteins.

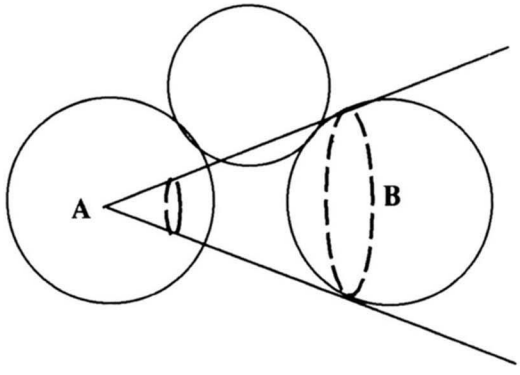

Fig. 1. Definition of shared surface area.

Potential function: (b) Hydrogen bonding and unsatisfied polar burials

Two observations suggest that a consequence of the tendency of nonpolar atoms to cluster is that many backbone polar atoms must also be buried, and buried polar atoms in the core will tend to cluster together. (1) About 66% of backbone polar groups (amides and carbonyl oxygens) are buried in the hydrophobic interiors of proteins, but they are almost always hydrogen bonded to a partner (McDonald & Thornton, 1994). (2) Model compound studies and calculations show that burying polar groups without hydrogen bonding in nonpolar media is energetically costly (Dill, 1990; Sharp et al., 1991) (see Fig. 2). Based on these observations, we assign an energy penalty to the burial of carbonyl or amide groups in the core that are not hydrogen bonded.

To determine when amides and carbonyls are hydrogen bonded, we use the following criteria from crystal data (Baker & Hubbard, 1984; Taylor & Kennard, 1984; Legon & Millen, 1987; McDonald & Thornton, 1994) and ab initio calculations (Del Bene, 1975): (1) linearity: the donor–H–acceptor angle is $\geq 90°$; (2) maximum length: the distance between H and acceptor is ≤ 2.5 Å; (3) lone pair plane: the hydrogen must lie within a dihedral angle of $\pm 60°$ of the lone pair plane. As an indirect consequence of nonpolar collapse, polar groups can be pushed to protein surfaces where they can form hydrogen bonds with water. Because those interactions are implicit in the nonpolar energy term, we include in the polar energy term only unfavorable polar interactions in the protein interior. This term has two parts. First, we count the number of "stand-alone" buried polar groups and give each an energy penalty of 1.5 kcal/mol, based on assuming that a hydrogen bond contributes around −3 kcal/mol. Second, "conflicting" buried polar pairs, i.e., donor/donor or acceptor/acceptor pairings in otherwise hydrogen bonding geometries, must be energetically costly. We assign conflicts an energy penalty of 1.5 kcal/mol. For k stand-alone interior polar groups and p pairs of polar conflicts, the energy cost is $(k + 2p) \times 1.5$ kcal/mol.

Thus, Geocore uses a very simple energy function, given explicitly in the footnote of Table 3. Even though we use only a single parameter for the hydrophobic interaction, that interaction applies to each nonpolar united atom rather than to each amino acid, so the present treatment allows for the individuality of the 20 different amino acids.

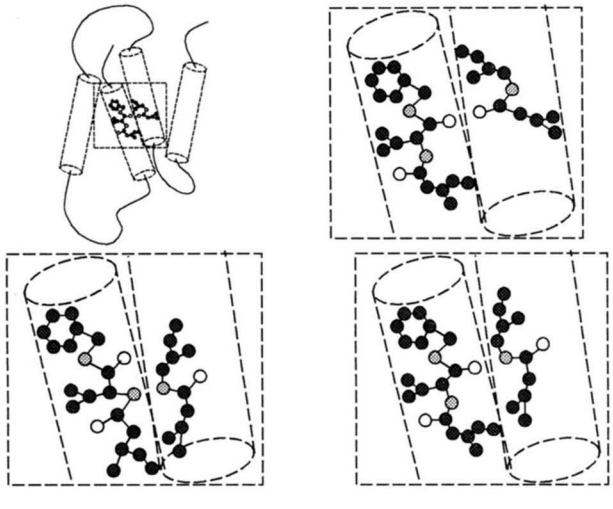

Fig. 2. United-atom model, indicating (top right) donor/donor conflict, (bottom left) acceptor/acceptor conflict, and (bottom right) a hydrogen bonded pair.

Constraint-based exhaustive searching

Monte Carlo or molecular dynamics methods sample conformational space sparsely. Exhaustive enumeration of conformational space has normally been considered computationally prohibitive for protein folding, but it offers the deepest test of potential energy functions. Here we describe a new method: "constraint-based exhaustive searching." Found in lattice models to be up to 37 orders of magnitude faster than brute force searching (Yue & Dill, 1995), constraint-based exhaustive searching is a branch-and-bound method that guarantees that all globally and near-globally optimal conformations will be found, while neglecting less important conformations.

Constraint-based exhaustive searching constructs conformations by sequential addition of residues in depth-first order (Aho et al., 1974). On the search tree, the nodes represent each added amino acid, and the different branches are the ϕ, ψ choices. When all the monomers are added or a dead end is reached, it backtracks. A complete traversal of such a tree will be an exhaustive search of the discrete set of rotational isomers. Geocore performs a complete search, subject to the two constraints that (1) no steric overlap is permitted and (2) the chain must be compact enough to lead to a near-maximal number of nonpolar contacts. Part of the basis for the search speed of the present method, compared to brute force exhaustive searching, is a rectangular solid boundary that tightly encloses the growing conformation. An earlier lattice implementation of the method (Yue & Dill, 1993, 1995) can find a boundary that is tight enough to be useful for pruning the search tree but is loose enough to guarantee retaining the globally optimal solution. In our current off-lattice implementation, which is presently somewhat cruder than the earlier lattice version, the bounded region has a volume typically about 60% higher than the native structure.

We treat side chain rotamers as follows. Each side chain dihedral angle starts in its most common rotameric position, according to its PDB distribution. If the value of the dihedral angle

causes a steric conflict, then the second most likely value for the dihedral angle will be tried. In this way, rotamer freedom is not neglected in the backbone search, but the computer cost is not prohibitive.

The following simple estimate shows that the drive to form hydrogen bonds among buried polar groups can be a strong constraint for pruning the conformational search tree. A consequence of the clustering of nonpolar atoms is that buried polar atoms will also cluster, and this pruning can be treated in a search strategy (Yue & Dill, 1995). Now we consider the further pruning involved in the formation of hydrogen bonds. Neglecting chain connectivity, if there are n main chain donors and n acceptors, then there are $n_1 = n!$ possible donor/acceptor pairings. But if *any* two polar groups could pair, then the number of combinations would be

$$n_2 = C(2n, 2)C(2n-2, 2)\cdots C(2(n-k), 2) \times C(2(n-k-1), 2)\cdots C(2, 2) = \prod_{k=1}^{n} (2k-1).$$

Because $(n_2/n_1) > 2^{n-1}$, only a very small fraction of conformations having good hydrophobic cores will also have proper donor/acceptor pairings. This may contribute to the uniqueness of protein structures.

Results

We ran Geocore to seek the low-energy conformations of four short proteins: crambin (1CRN), avian pancreatic polypeptide (1PPT), melittin (2MLT), and apamin (sequence: CNCKAPET ALCARRCQQH). This set of proteins was chosen simply because it represents, as far as we know, all the proteins of known structure that are within the chain length that can currently be explored by our method. Geocore does not produce a single lowest-energy conformation that resembles the true native protein at high resolution. Rather, Geocore produces a relatively small ensemble of conformations, among which some bear good resemblance to the true native structure. Our larger objective here is not to find an optimal energy function that can select a single "right answer" for this small set of proteins, but rather to find a "simplest" energy function "filter" that may be useful ultimately on a wider set of proteins.

Table 2 shows a typical run for 1PPT, in which 1.9 billion nodes are visited and 0.19 billion conformations are constructed. The runtime is proportional to the number of nodes visited. On the average, 22,000 residue nodes can be searched in a minute on a Sparc 10 workstation. For the run shown for 1PPT, approximately 8,217 conformations have low energy, defined here as being no more than 16 kcal/mol above the global optimum. The conformations are found to be reasonable by two criteria: (1) solvent-accessible nonpolar surface areas of representative conformations calculated with the ACCESS program (Hubbard, 1991) are generally consistent with our counts of HH contacts;

Table 2. *Test run for 1PPT with elongated conformation boundary*[a]

Branch	t_{HH}	H	B	D	Clusters	C	N	RMSD$_{min}$
00	0.0	0	–	0		0	6	–
010	294	17	4	130	39	18,284,976	178,572,575	4.49
011	0.0	0	–	0	0	6	–	
012	290.9	33	0	954	239	7,269,757	82,036,747	4.52
013	299.1	16	1	159	55	18,121,269	172,003,296	6.0
02	279.6	20	2	2,311	1,150	15,613,685	167,665,002	4.3
03	297	24	4	105	32	37,356,998	398,636,852	4.5
1100	294.5	36	0	1,314	391	8,108,538	75,166,230	4.79
11010	294.1	28	0	480	190	7,800,000	80,000,000	5.6
11011	0.0	0	–	0	0	6	–	
11012	291.0	32	0	411	81	2,359,839	22,278,363	5.3
11013	298.7	27	0	355	200	3,989,473	36,125,063	5.2
111	0.0	0	0	–	0	0	4	–
112	292.4	7	0	685	130	18,378,400	170,745,454	4.77
113	297.6	17	2	554	56	16,414,787	156,053,531	5.16
12	283.4	19	2	527	142	9,770,495	115,898,918	4.43
13	293	21	4	232		25,047,606	283,863,310	4.33
2	0.0	0	–	0	0	6	–	

[a] Data for 1PPT were obtained with steric conflict allowance of 0.2 Å and a conformational boundary of 33.5 × 16.5 × 14.5 Å. Because a full search was not possible for four ϕ, ψ choices for every amino acid, we chose arbitrarily to allow four ϕ, ψ choices for the first 12 residues and three for the remainder, because this choice introduces no more bias than any other and still allows more flexibility than 3 choices uniformly along the chain. In this case, the ϕ, ψ angles that comprise the choice set are the most frequent four (or three) in the PDB. "Branch" indicates the position on the search tree, in our numbering system; t_{HH} is the maximum number of HH contacts; D is degeneracy, i.e., the number of conformations with energies that are the same as or close to the minimum. C is the total number of constructed conformations, and N is the total number of (residue) nodes visited. The optimal number of stand-alone polar groups and polar groups in conflict is B, and the optimal number of main chain hydrogen bonded groups is H. "Clusters" indicates the number of clusters for a given branch. RMSD$_{min}$ is the minimum Cα RMSD (in Å) for the low-energy conformations found in the branch relative to the native 1PPT conformation.

(2) the nonbond and hydrogen bonding energies of low-energy conformations are comparable with the native structures when calculated using default AMBER forcefield parameters. We have not used subsequent energy minimizations or other refinements.

To compare the computed structures with the native proteins, we compute RMS deviations (RMSD) of the Cα coordinates. To avoid storing too many conformations, particularly those that differ by only a few bonds, we keep all the energies but not all the conformations. We record the coordinates for a maximum of 400 low-energy conformations per branch. Beyond 400, we use a method designed to uniformly record the conformations. For example, the method avoids recording two conformations differing by only a single residue when the total number of low-energy conformations exceeds 400.

Figures 3 and 4 show the distribution of energies versus RMSDs for 1PPT and 1CRN. Because many conformations are geometrically similar, we define a set of conformations as being in the same "cluster" if their pairwise RMSDs are less or equal to 3.5 Å. Compared with the number of low-energy conformations, the number of "clusters" is significantly reduced, as shown in Table 2. On this basis, the most native-like computed conformation of 1PPT (as determined by RMSD) is within the 100 lowest-energy clusters and for 1CRN is within the 200 lowest-energy clusters.

In our model, neither energy term alone is sufficient to give native-like structures. Table 3 shows that structures with too many HH contacts are too compact and restrict hydrogen bonding, whereas structures that only have good polar energies are too open (see Figs. 5 and 6). With this simple energy function, the true native structure is always better than any enumerated conformation by at least 3–10%. This is evidence that the discrepancies between our best model conformations and the true native structures of these four proteins are due to our restricted conformational choices, not to flaws in the energy function.

Native-like low-energy conformations are shown in Figure 7 compared with the known native structures. The computed structures shown are not those having a global minimum of free energy. Rather, among the 100 or so lowest-energy conforma-

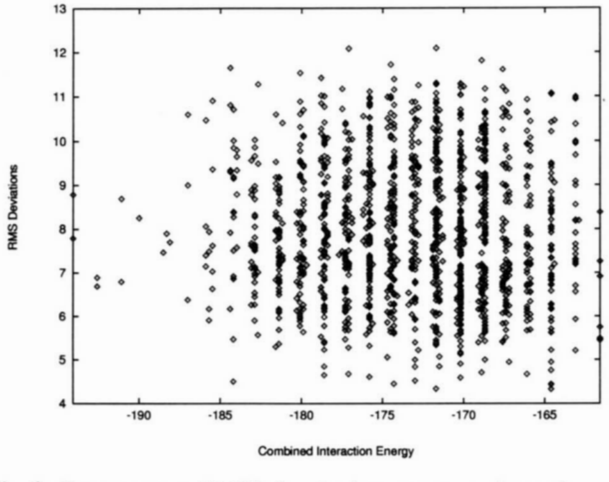

Fig. 3. Energy versus RMSD for the low-energy conformations of 1PPT. Each point is a cluster of conformations. The native state, which has an energy of −196.6 (Table 3), is not shown here.

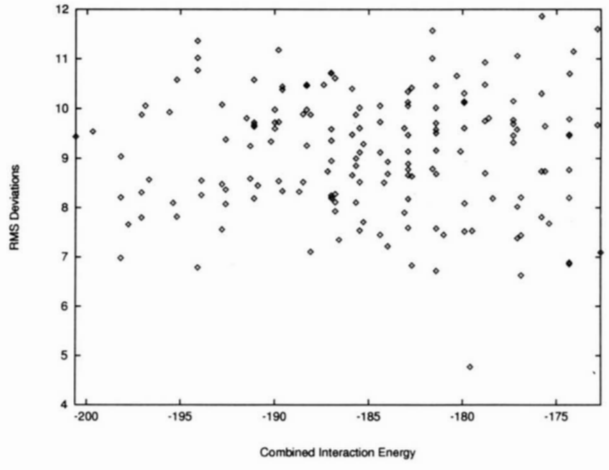

Fig. 4. Energy versus RMSD for the low-energy conformations of 1CRN. The native state, which has an energy of −233.7 (Table 3), is not shown here.

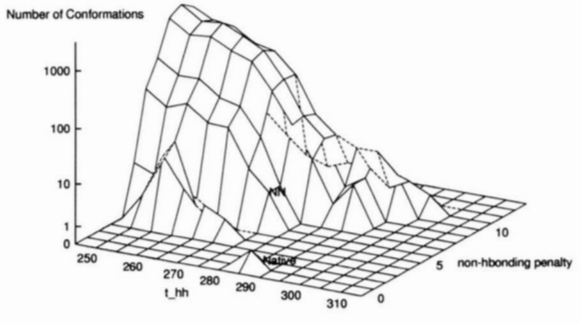

Fig. 5. Energy distribution of lowest-energy states for 1PPT. Because Geocore uses only two energy quantities (the number of HH contacts [t_{HH}] and the number of unsatisfied polar burials [see footnote of Table 3]), the set of "best" conformations can be described without resorting to any particular choice of energy parameters, as was done in Table 3. We show the best and near-best values for the two energy parameters, given as the x, y coordinates here. Heights indicate the number of conformations found, on a log scale. The small peak at $(x, y) = (283, 1)$ is the native conformation. The point labeled "NN" corresponds to 1PPT #2 in Table 3. The valley separating the native from predicted structures implies that the ϕ, ψ angle choices are too few in the present search to represent the native structure more accurately.

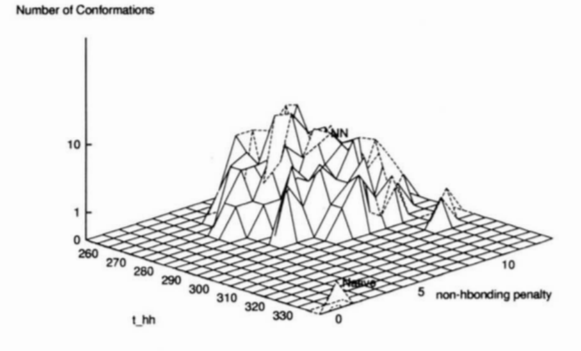

Fig. 6. Energy distribution of lowest-energy states for 1CRN.

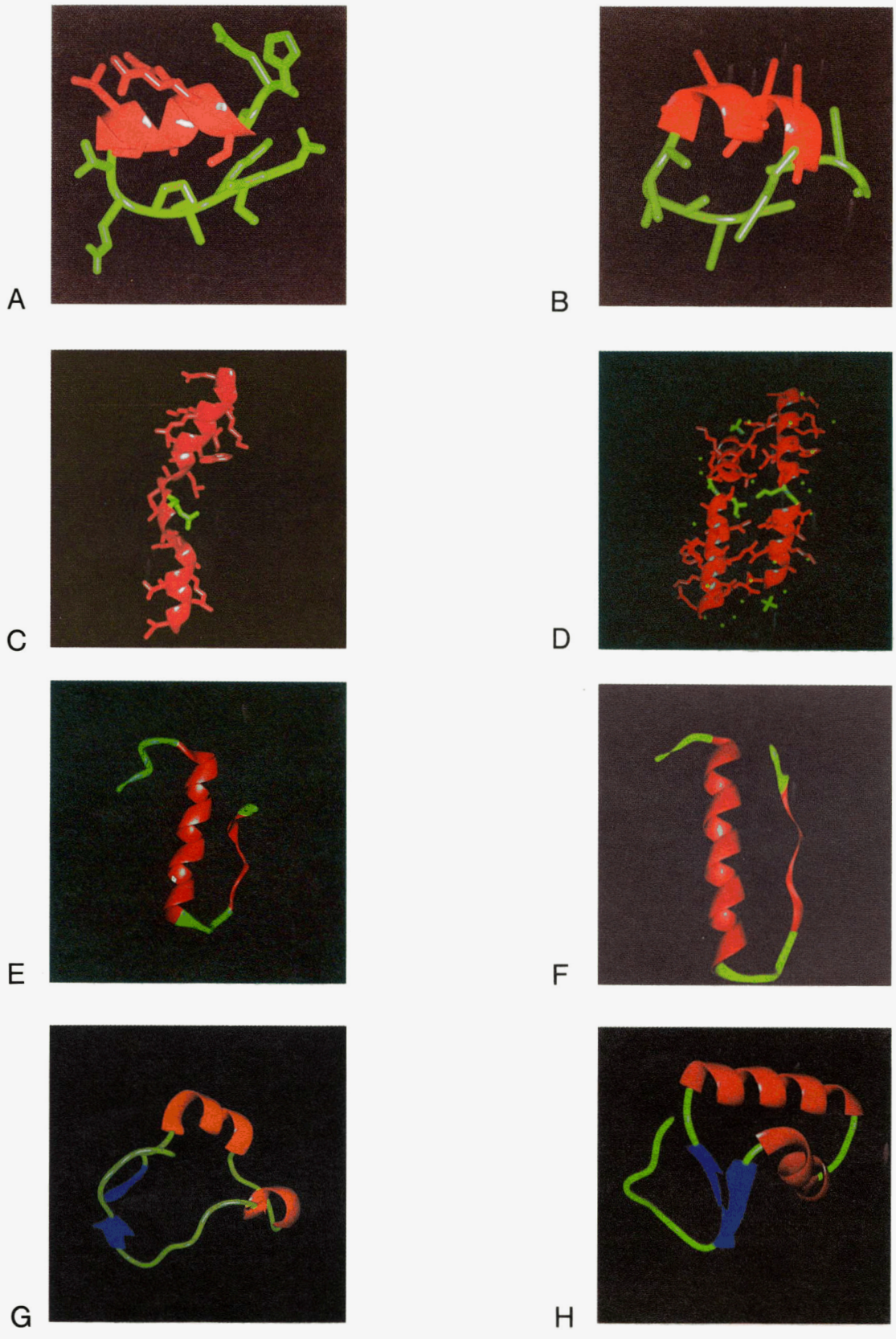

Fig. 7. Structure comparisons of the native state with the most similar structures from the low-energy ensemble. Side chain atoms are shown for 2MLT and apamin. Crambin: (**A**) native, (**B**) predicted (1CRN #2 in Table 3). 1PPT: (**C**) native, (**D**) predicted (1PPT #2 in Table 3). 2MLT: (**E**) native, (**F**) predicted, RMSD = 2.85 Å. Apamin: (**G**) native (only Cα coordinates are available), (**H**) predicted, RMSD = 2.6 Å.

Table 3. *Energies of selected conformations of 1CRN and 1PPT*[a]

	L	t_{HH}	N_{HBMC}	B_{NH}	B_{CO}	C_m	E	RMSD
1CRN Native	46	336	56	0	1	0	−233.7	0
1CRN #1	46	330	16	7	5	4	−201.0	9.0
1CRN #2	46	278	25	5	3	1	−179.6	4.6
1PPT Native	36	283	36	1	0	0	−196.6	0
1PPT #1	36	290	3	4	3	1	−189.5	9
1PPT #2	36	274	31	2	1	0	−187.3	4.4

[a] L is the chain length, t_{HH} is the number of HH contacts, N_{HBMC} is the number of hydrogen bonded main chain polar groups, B_{NH} and B_{CO} are the numbers of buried stand-alone donors and acceptors, and C_m is the number of main chain polar conflicts. The total energy E is

$$E = -0.7t_{HH} + 1.5(B_{NH} + B_{CO} + 2C_m).$$

The term in parentheses is the penalty in Figure 5 for donors and acceptors that avoid hydrogen bonding. The two energy quantities −0.7 and 1.5 have not been optimized. The disulfide bond energy is not included in the above expression. Instead, the requirement that sulfur atoms should form disulfide bonds is used as a loose constraint in the conformational search for 1CRN. It is loose in that it only requires that a sulfur atom participates in a disulfide bond, but no a priori (e.g., native) pairing is given. 1CRN #2 is the most native-like structure in a complete search (lowest RMSD relative to native). 1PPT #2 is the most native-like structure in a complete search when an average of 3.2 ϕ, ψ isomers are explored per residue.

tions, we show here the single conformation that has the best RMSD relative to the native structure. We show these figures only to indicate the degree to which Geocore retains a native-like structure in a small ensemble. We do not feel a more detailed analysis is currently warranted. Many of the 100 low-energy structures have native-like features in common. The main result here is that despite the extreme simplicity of the energy function (see also Srinivasan & Rose, 1995; Sun et al., 1995), it is adequate for discriminating native from non-native structures in a much more extensive conformational search than has been possible before.

Conclusions

We propose a very simple energy function, based on the burial of polar and nonpolar amino acids, that can recognize native structures of proteins. The conformational search explores low-energy states more extensively than previous methods. Native-like structures of four small proteins are found as a compromise between the tendency to form very good hydrophobic cores and to avoid burying unsatisfied or conflicting carbonyl and amide groups. We believe the main novelty of the present work for computational protein folding is in its greater simplicity—fewer parameters and they are physical, based on hydrophobic and hydrogen bond interactions. This work suggests that there may be practical strategies for extensive conformational searching that find stable states of proteins through use of very simple energy functions. As a practical matter, the two main limitations at the moment are the restricted flexibility dictated by the discrete set of ϕ, ψ angles, which limits the accuracy with which the native protein can be represented by the computable conformations and the search speed for reaching longer chain lengths.

Acknowledgments

We thank the ONR for financial support. Kaizhi Yue thanks Dr. Yuzhong Wang for helpful discussions.

References

Aho A, Hopcroft J, Ullman J. 1974. *The design and analysis of computer algorithms*. New York: Addison-Wesley.
Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science 181*:223-230.
Baker EN, Hubbard RE. 1984. Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol 44*:97-197.
Boczko EM, Brooks CL. 1995. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science 269*:393-396.
Cantor CR, Schimmel PR. 1980. *Biophysical chemistry*. New York: Freeman.
Covell DG. 1992. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins Struct Funct Genet 14*:409-420.
Covell DG. 1994. Lattice model simulations of polypeptide chain folding. *J Mol Biol 235*:1032-1043.
De la Cruz X, Reverter J, Fita I. 1992. Representation of noncovalent interactions in protein structures. *J Mol Graphics 10*:96-110.
Del Bene J. 1975. Molecular orbital theory of the hydrogen bond. XII. *J Chem Phys 62*:1961-1970.
Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133-7155.
Dill KA, Bromberg S, Yue K, Fiebig KM, Thomas PD, Chan HS. 1995. Principles of protein folding—A perspective from simple exact models. *Protein Sci 4*:561-602.
Eisenberg D, Weiss RM, Terwillinger TC. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci USA 81*.
Flory PJ. 1969. *Statistical mechanics of chain molecules*. New York: Interscience Publishers.
Hinds D, Levitt M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol 243*:668-682.
Hubbard S. 1991. *ACCESS*. London: UCL.
Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct Funct Genet 18*:338-352.
Kuntz ID, Crippen GM, Kollman PA, Kimelman D. 1976. Calculation of protein tertiary structure. *J Mol Biol 106*:983-994.
Lee BK, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379-400.
Legon AC, Millen DJ. 1987. Directional character, strength, and nature of the hydrogen bond in gas phase dimers. *Acc Chem Res 20*:39-45.
Levitt M, Warshel A. 1975. Computer simulation of protein folding. *Nature 253*:694-698.
McDonald I, Thornton J. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol 238*:777-793.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecule 18*:534-552.
Monge A, Friesner RA, Honig B. 1994. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc Natl Acad Sci USA 91*:5027-5029.

Novotny J, Rashin AA, Bruccoleri RE. 1988. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct Funct Genet 4*:19–30.

Sharp K, Nicholls A, Fine RF, Honig B. 1991. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science 252*: 106.

Sippl M, Hendlich M, Lackner P. 1992. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin beta 4. *Protein Sci 1*:625–640.

Skolnick J, Kolinski A. 1990. Simulations of the folding of a globular protein. *Science 250*:1121–1125.

Srinivasan R, Rose G. 1995. LINUS—A hierarchic procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.

Sun S, Thomas PD, Dill KA. 1995. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng*. Forthcoming.

Taylor R, Kennard O. 1984. Hydrogen bond geometry in organic crystals. *Acc Chem Res 17*:320–326.

Vajda S, Jafri MS, Sezerman OU, DeLisi C. 1993. Necessary conditions for avoiding incorrect polypeptide folds in conformational search by energy minimization. *Biopolymers 33*:173–192.

Wallqvist A, Ullner M. 1994. A simplified amino acid potential for use in structure predictions of proteins. *Proteins Struct Funct Genet 18*:267–280.

Wilson C, Doniach S. 1989. A computer model to dynamically simulate protein folding—Studies with crambin. *Proteins Struct Funct Genet 6*:193–209.

Yue K, Dill KA. 1993. Sequence structure relationship of proteins and copolymers. *Phys Rev E 48*:2267–2278.

Yue K, Dill KA. 1995. Forces of tertiary structural organization of globular proteins. *Proc Natl Acad Sci USA 92*:146–150.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA 92*:325–329.