Protein Science (1999), *8*:716–721. Cambridge University Press. Printed in the USA.

Predicting the structures of 18 peptides using Geocore

KOHKI ISHIKAWA,[1] KAIZHI YUE,[2] AND KEN A. DILL[2]

[1]Central Research Laboratories, Ajinomoto Co., 1-1 Suzuki-cho, Kawasaki 210, Japan
[2]Department of Pharmaceutical Chemistry, University of California at San Francisco, Box 1204, San Francisco, California 94143

(RECEIVED June 18, 1998; ACCEPTED November 28, 1998)

Abstract

We describe an extensive test of Geocore, an ab initio peptide folding algorithm. We studied 18 short molecules for which there are structures in the Protein Data Bank; chains are up to 31 monomers long. Except for the very shortest peptides, an extremely simple energy function is sufficient to discriminate the true native state from more than 10^8 lowest energy conformations that are searched explicitly for each peptide. A high incidence of native-like structures is found within the best few hundred conformations generated by Geocore for each amino acid sequence. Predictions improve when the number of discrete ϕ/ψ choices is increased.

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

We describe an extensive test of a simple computer algorithm, called Geocore, for predicting the three-dimensional structures of peptides from their amino acid sequences (Yue & Dill, 1996). Geocore differs from other ab initio protein folding algorithms (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; Srinivasan & Rose, 1995) in several respects. First, Geocore is intended as a *filtering* algorithm, rather than a *folding* algorithm. It aims to find a small ensemble of conformations, within which are native-like structures, rather than to find the single best conformation. While a folding algorithm is obviously more desirable in the long run than a filtering algorithm, we believe that simplified models, at least in their present state of development, may not be sufficiently good to discriminate subtle differences (Dill, 1997). If the ultimate aim is to be predictive for the broadest possible range of protein structures, then overmodeling to force a few sequences to fold to their single native conformations may be counterproductive for ultimately predicting the folded structures of other proteins. Hence, in recognition of the limitations of simple models, our more modest goal here is just to develop a filtering algorithm.

Second, Geocore is unique in its extensive conformational search strategy. While this limits the method to short chains, i.e., peptides shorter than about 30 amino acids at the present time, it has the advantage of providing a deep test of the energy function. When other folding methods fail, it is often unclear whether the problem is a poor energy function or an incomplete search of it. In Geocore, because the wide coverage of the conformational space, failures can be attributed unambiguously to the energy model. The large ensemble of conformations generated by Geocore, often numbering over billions, provides the data to evaluate what is wrong with the potential function. We believe this is an essential step toward building folding algorithms that can be refined and improved.

Third, among folding algorithms, Geocore has arguably one of the simplest energy functions, with relatively few parameters. We find here that when we add ϕ/ψ choices, taken from the study of PA Karplus (Karplus, 1996), the predictions of Geocore are improved. This implies that the energy function is not limiting, even in this simple model.

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

The Geocore algorithm

Here we summarize the method; details are given in Yue and Dill (1996). Each amino acid is represented at the united-atom level, with polar hydrogens included explicitly, for the purpose of hydrogen bonding. United atoms include methylene groups, amide groups, hydroxyl groups, etc. Each (united) atom is a hard sphere with its appropriate van der Waals (vdW) radius, but with a tolerance that allows two atoms to overlap by 0.2 to 0.5 Å. Backbone conformations are represented by discrete sets of torsion angles (ϕ/ψ). Standard values are used for bond lengths and bond angles. The user has the option to specify the value of steric tolerance and the values of ϕ/ψ angles. The default numbers and values of the ϕ/ψ angle preferences for each amino acid are extracted from the Protein Data Bank (PDB) (Yue & Dill, 1996).

The Geocore energy function has two terms, hydrophobic interaction and hydrogen-bond energy (Yue & Dill, 1996). Geocore seeks conformations with minimal nonpolar exposure to the sol-

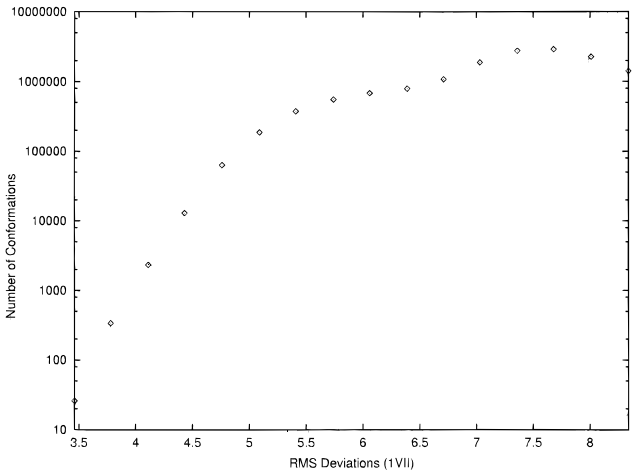

Fig. 1. Distributions of conformations by RMSDs for 1VII.

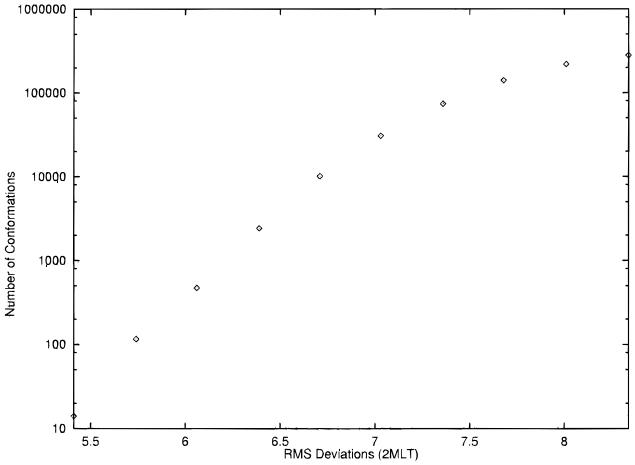

Fig. 2. Distributions of conformations by RMSDs for 2MLT.

vent. This is implemented by finding maximal pairwise shared nonpolar surface areas among nonpolar atoms, called "HH contacts." When two carbon united atoms contact, the energy is -0.7 kcal/mol. Since the drive for polar groups to hydrogen-bond can be satisfied by either bonding with water or with polar groups in the protein, Geocore assigns an energy penalty to the burial of carbonyl or amide groups in the core that are not hydrogen bonded. Each buried polar group that is not H-bonded has an energy penalty of 1.5 kcal/mol.

Geocore constructs conformations by adding one residue at a time to a growing chain. By adding residues with different ϕ/ψ angles defined by the chain representation, Geocore exhaustively considers all the conformations, even without explicitly evaluating them. A branch and bound method is used that guarantees that all globally optimal and near-globally optimal conformations will be found, while neglecting less important conformations. The search is done in depth-first order (Aho et al., 1974). On the search tree, the nodes represent added amino acids and the different branches are the ϕ/ψ choices. When the full chain length or a dead end is reached, the search backtracks. Geocore performs a complete search, subject to the two constraints that steric overlap is not permitted (in excess of the tolerance criterion), and that the chain must be compact enough to lead to a near maximal number of nonpolar contacts. Geocore gives the user the option to specify possible bounds on the shape of allowed conformations. Geocore is written in C and runs on most hardware platforms. The work described here was performed mainly on a Pentium-Pro-based personal computer.

For each amino acid sequence, for each run of Geocore, we retain only approximately 400 "best" conformations, as defined by either of two criteria. (1) We keep 400 conformations that are

Table 1. *Proteins tested*

Protein	Sequence	Description
1WBR	QAERMSQIKRLLSEKKT	Human CD4 receptor peptide
1PAO	ACKSTQDPMFTPKGCDN	PAO Pilin Trans peptide
1EDP	CSCSSLMDKECVYFCHL	Endothelin I
1FGE	CEAPEGYILDDGFICTDIDE	Thrombomodulin
1TER	ALCNCNRIIIPHMCWKKCGKK	Tertiapin
1OMG	CKGKGAKCSRLMYDCCTGSCRSGKC	Omega-conotoxin
1ANS	RSCCPCYWGGCPWGQNCYPEGCSGPKV	Neurontoxin III
2ETI	GCPRILMRCKQDSDCLAGCVCGPNGFCG	Trypsin inhibitor
1KAL	SWPVCTRNGLPVCGETCVGGTCNTPGCTC	Kalata B1
1MMC	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR	AC-AMP2
1SCY	AFCNLRMCQLSCRSLGLLGKCIGDKCECVKH	Scyllatoxin
1DEP	RSPDFRKAFKRLLCF	Beta-adrenoreceptor peptide
1ALE	ALDKLKEFGNTLEDKARE	Apolipoprotein C-I, residues 7–24
1ODR	YSDELRQRLAARLEALKENG	Human APOA-I peptide
1BTR	VLAAVIFIYFAALSPAITFG	Human manc 3, synthetic peptide
1SOL	KHVVPNEVVVQRLFQVKGRR	PIP2 and F-a. of gelsolin
1FAC	TRYLRIHPQSWVHQIALRMEV	Coagulation factor VIII
1PEI	VEEKSIDLIQKWEEKSREFIGS	CTP phisphichholic peptide

[a]Proteins starting from 1DEP are not water-soluble.

Table 2. *Disulfide bonds in water-soluble proteins*

Protein	Sequence	Disulfide bonds
1WBR	QAERMSQIKRLLSEKKT	
1PAO	ACKSTQDPMFTPKGCDN	2–15
1EDP	CSCSSLMDKECVYFCHL	3–11, 1–15
1FGE	CEAPEGYILDDGFICTDIDE	1–15
1TER	ALCNCNRIIIPHMCWKKCGKK	3–14, 5–18
1OMG	CKGKGAKCSRLMYDCCTGSCRSGKC	1–16, 8–20, 15–25
1ANS	RSCCPCYWGGCPWGQNCYPEGCSGPKV	4–11, 3–17, 6–22, III
2ETI	GCPRILMRCKQDSDCLAGCVCGPNGFCG	2–19, 9–21, 15–27
1KAL	SWPVCTRNGLPVCGETCVGGTCNTPGCTC	5–22, 13–27, 17–29
1MMC	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR	4–15, 9–21, 14–28
1SCY	AFCNLRMCQLSCRSLGLLGKCIGDKCECVKH	3–21, 8–26, 12–28

among the lowest in energy (energy-based criterion). (2) We keep the 400 conformations that are among the lowest in root-mean-square deviation (RMSD) relative to the true native structure, as defined by the PDB coordinates (geometry-based criterion). The latter is just to test the adequacy of the chain representation. For the purpose of deciding which 400 to keep, we use a sampling algorithm that skips geometrically similar conformation to ensure a representative ensemble in the program output.

The Geocore program can make use of disulfide bond information. The user can specify which cysteine residues form disulfide bridges, or can specify only that some form, and let the program find them. Specifying the disulfide bonds biases the search and speeds it up. To test this bias, we have compared the RMSDs of conformations generated with and without assumptions of disulfide bonds for endothelin (1EDN), a 21mer peptide. In the runs without the disulfide bond assumption, we found a minimum RMSD from the native structure of 3.8 Å and an average RMSD of 5.94 Å, with a standard deviation of 0.95 Å. With the disulfide bonds assumed as a constraint, we found a minimum deviation of 3.0 Å and an average RMSD of 4.6 Å, with standard deviation of 0.65 Å. We note that for a conformational space in which each residue has four ϕ/ψ choices, the total number of Geocore generated compact conformations for 1EDN is approximately 51 million. The numbers of conformations with RMSDs of 6.0 or less, 4.6 Å or less, and 3.8 Å or less are 28 million, 5.6 million, and 0.8 million, respectively. Figures 1 and 2 show the numbers of conformations sampled by Geocore as a function of RMSD from the native structure for 1VII (villin head piece) and 2MLT (melittin). This shows that most conformations deviate by 7–8 Å, and very few are native-like. It indicates that native-like structures that are being found by Geocore are not due to some property of the constraints or the search, but are due to the energy function, 18 peptides are tested in our study. Tables 1 and 2 show the proteins tested and disulfide bonds in water-soluble proteins, respectively.

Results

Comparing the first two columns of Table 3 shows that for 8 of the 11 water-soluble proteins, the true native structure has a lower value of the Geocore energy function than the lowest energy structure computed by Geocore. This means that the Geocore energy function is perfectly adequate for the job it is supposed to perform: It can recognize real native structures, and can distinguish them from poorer conformations. (By "recognize" (Maiorov & Crippen, 1992), we mean that the energy function reports the native structure to have lower energy than the alternatives.) The three exceptions, 1WBR, 1PAO, and 1OMG, are relatively small and have too little hydrophobic core for hydrophobicity to dominate the energy. According to these results, native structures, even in peptides, are substantially driven by hydrophobic interactions.

A test of how well the energy function can discriminate the true native structure from the most nearly native structures the model

Table 3. *Energies (kcal/mol) of native structures, energy-based most native-like conformations, and geometry-based most native-like conformations*[a]

Protein	Energy of native structure (kcal/mol)	Energy of conformations of energy-based search (kcal/mol)	Energy of conformations of geometry-based search (kcal/mol)
1WBR	−72.2	−101.5	−61.0
1PAO	−62.6	−84.3	−66.8
1EDP	−69.2	−68.4	−59.6
1FGE	−120.9	−107.6	−91.2
1TER	−152.2	−97.9	−75.9
1OMG	−106.4	−108.3	−87.8
1ANS	−158.4	−120.0	−106.2
2ETI	−122.8	−103.3	−103.3
1KAL	−161.2	−109.3	−99.4
1MMC	−157.3	−103.5	−87.1
1SCY	−166.7	−137.30	−121.9
1DEP	−83.0	−85.8	−64.5
1ALE	−94.3	−80.4	−76.2
1ODR	−86.2	−93.5	−82.9
1BTR	−94.9	−91.2	−76.3
1SOL	−71.3	−95.0	−94.0
1FAC	−89.6	−114.6	−88.9
1PEI	−105.5	−108.0	−99.3

[a] Here, a conformation is judged "most native-like" if its RMSD from the native structure is minimal among all possible conformations. Energy-based most native-like conformations are chosen from the pool of low energy conformations. Geometry-based native-like conformations are found from the entire conformational space. Proteins starting from 1DEP are not water-soluble.

Table 4. *RMSD values of C-α atoms between a native structure and an energy/geometry-based most native-like conformation*[a]

Protein	RMSDs of conformations of geometry-based search (Å)	RMSDs of conformations of energy-based search (Å)
1WBR	1.749	3.010
1PAO	1.847	2.900
1EDP	2.183	2.461
1FGE	2.480	2.992
1TER	2.952	3.464
1OMG	2.608	3.904
1ANS	4.221	4.880
2ETI	4.586	4.586
1KAL	4.552	4.835
1MMC	4.680	5.508
1SCY	4.028	5.090
1DEP	1.000	2.113
1ALE	0.913	2.769
1ODR	1.500	3.233
1BTR	2.098	4.000
1SOL	1.360	3.138
1FAC	1.826	3.039
1PEI	0.953	3.133

[a]Proteins starting from 1DEP are not water-soluble.

can produce is given by comparing columns one and three of Table 3. In 16 of the 18 molecules, the energy function correctly distinguishes the true native structure from the most native-like structure that the model can produce. This means the current limitation is not a poor energy function, but the inability of the chain to reach a better conformation, due to ϕ/ψ limitations. One of the two failures is 1SOL, which is not water-soluble. The other is 1PAO, which also failed the test described above, and is small. Hence, with these few exceptions, the energy function is an adequate discriminator of native from nonnative structures.

Column one of Table 4 shows the geometric limitations of the model. Shown are the RMSD values between the true native structure and the best structures the model can produce. Remarkably, the water-insoluble proteins are more accurately captured by the canonical ϕ/ψ values in the model chain representation than the water-soluble proteins. Errors are generally larger in larger peptides. Comparison of column two to column one shows that Geocore's lowest energy conformations are usually not much worse than Geocore's best geometric structures. Said differently, the present main limitation of Geocore is the chain representation and the discreteness of the ϕ/ψ options.

To test this, we performed a limited test on the few peptides that were short enough that we could increase the computational search from four ϕ/ψ options to five. Table 5 shows that when the number of options is five, the Karplus values for the most probable ϕ/ψ's are generally an improvement over our original default values. Table 6 shows the result of using four ϕ/ψ options. Comparison of the two tables shows, not surprisingly, that using five ϕ/ψ options rather than four, improves the performance of the model with respect to the true native structures. This is a further indication that computer time is a greater limitation at present for this algorithm than any weakness in the physical model.

The limitations of the Geocore model can be seen in Figure 3. Hydrophobic interactions compete with the tendencies toward helical structures and the helices of water-insoluble proteins are not well predicted. On the other hand, Geocore was not intended for water-insoluble proteins; we included them here because we were

Table 5. *Comparison of RMSD values between default and Karplus ϕ/ψ values, five ϕ/ψ choices*

Protein	Chain length	Default ϕ/ψ geometry-based search (Å)	Default ϕ/ψ energy-based search (Å)	Karplus ϕ/ψ geometry-based search (Å)	Karplus ϕ/ψ energy-based search (Å)
1PAO	17	1.85	2.62	1.31	2.03
1EDP	17	2.16	2.44	1.46	2.10
1TER	21	2.92	3.44	2.30	2.98
1ANS	27	4.03	5.00	3.31	6.25
1DEP	15	0.86	2.67	0.97	2.30

Table 6. *Comparison of RMSD values between default and Karplus ϕ/ψ values, four ϕ/ψ choices*

Protein	Chain length	Default ϕ/ψ geometry-based search (Å)	Default ϕ/ψ energy-based search (Å)	Karplus ϕ/ψ geometry-based search (Å)	Karplus ϕ/ψ energy-based search (Å)
1PAO	17	1.85	2.90	1.61	2.41
1EDP	17	2.18	2.46	1.80	2.65
1TER	21	2.91	3.32	2.93	3.55
1ANS	27	4.25	4.92	4.08	4.93
1DEP	15	1.00	2.11	1.27	3.01

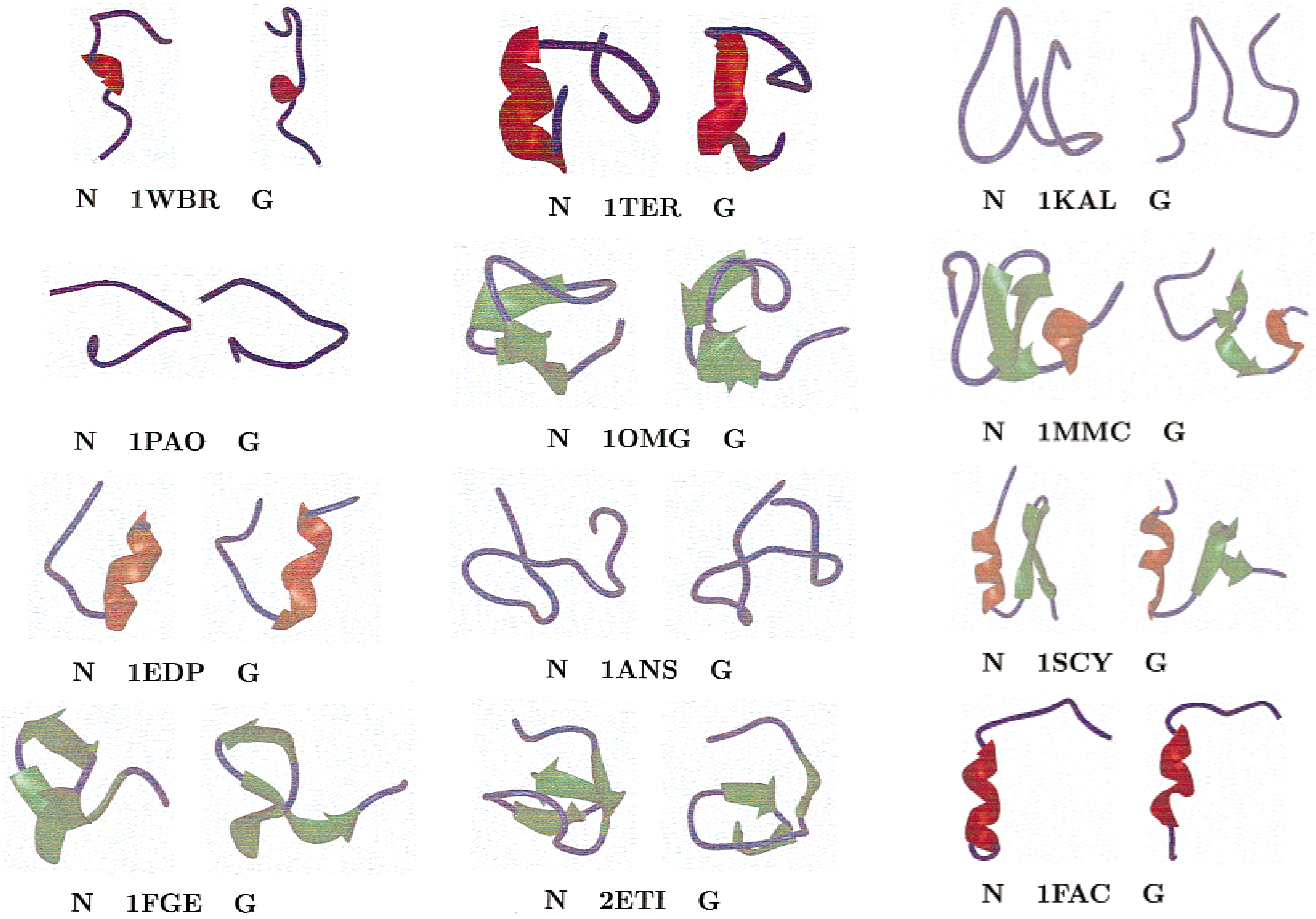

Fig. 3. Ribbon diagrams of the native structure (N) from the PDB and Geocore-generated low energy structure (G) for each of the water-soluble proteins studied and 1FAC, one of the water-insoluble proteins. The figure labeled (G) in each pair of structures represents not the very lowest energy conformation but the most similar structurally from the set of around 400 recorded lowest energy conformations.

interested to see what the algorithm would do with them. For water-soluble proteins, the most native-like energy-based conformation is not considerably different from the most native-like geometry-based conformation, but these are not always native-like. This is probably due to the limited number of ϕ/ψ choices in the conformational search.

Summary

We have tested an algorithm called Geocore on the prediction of the structures of 18 peptides from their amino acid sequences. Geocore uses a very simple energy function and a very complete conformational search method. The energy function has essentially two parameters, one for hydrophobic interactions and the other for the burial of polar groups, in addition to a few steric tolerance parameters and PDB-derived ϕ/ψ propensities. Despite its simplicity, the energy function is sufficient to distinguish native from a very extensive list of non-native conformations. The main limitation at the present time is the discreteness of the ϕ/ψ options and the computational limitations to chain lengths less than about 30–40 monomers. Compared to our earlier report on this algorithm (Yue & Dill, 1996), we find that the algorithm is improved either by using more ϕ/ψ options per amino acid, or by using the Karplus ϕ/ψ propensities.

Materials and methods

Choosing the test proteins

We chose 18 peptides to study, based on the following criteria: structures were known and available in the PDB; chain lengths were restricted to 22 amino acids when we applied no disulfide constraints or 31 when they were included; we eliminated molecules that crystallize as dimers or that involve prosthetic groups. To avoid bias, we otherwise took all peptides that meet those criteria. In all cases, structures were determined by NMR. Eleven of the 18 protein structures were determined in aqueous solution, while the other 7 were determined in the presence of detergent or organic solvents, because they are otherwise insoluble or adopt multiple conformations.

In the conformational search each residue has four ϕ/ψ choices, except glycines and the residues around glycines (two residues before and two after), which have one additional ϕ/ψ choices. The numbers of generated conformations and the run time are listed in Table 7.

Table 7. *Number of conformations and run time*

Protein	Number of conformations	Run time
1WBR	281317903	40h 1m
1PAO	382342	3h38m
1EDP	3542	39m
1FGE	26171948	20h34m
1TER	22954	32m
1OMG	104102171	77h53m
1ANS	293867	21m
2ETI	9060	115h19m
1KAL	295392	210h28m
1MMC	384024	26h46m
1SCY	16982	416h46m

Comparing predictions to native structures

Because the molecules we study are peptides and their structures are determined by NMR, there is no single true native structure. The PDB files contain multiple conformations. The energy of a native structure was determined by averaging the energies of all the conformations in the PDB file. Table 3 compares the energies: (1) the lowest energy conformation found by Geocore, (2) the energy computed with the Geocore energy function, averaged as noted above, for the native structure, and (3) the energy of the most native-like structure that is possible within the Geocore chain representation.

Table 4 shows the RMSDs of C-α coordinates between generated and native conformations. Figure 3 is the ribbon diagrams that show these comparisons.

The ϕ/ψ values of PA Karplus

To see if it is possible to improve the RMSD values relative to a native structure, another set of ϕ/ψ values were tested. PA Karplus (Karplus, 1996) reported ϕ/ψ-distributions from 70 diverse proteins refined at 1.7 Å or better. The ϕ/ψ values found in high-resolution crystal structures cluster around definite regions with fairly sharply defined borders on the edges of these regions in the Ramachandran plot. From the Karplus study, we assigned ϕ/ψ = 295/315, 90/0, 55/35, 75/190, 270/170, and 270/0 for Gly; 300/300 and 285/145 for Pro; 300/315 and 240/125 for Val and Ile; 300/320, 240/140, 280/140, 60/40, and 270/0 for other residues.

For residues other than Gly, Pro, Val, and Ile, our approach toward systematic improvement was to use five ϕ/ψ values to cover the Ramachandran plot, rather than the four ϕ/ψ choices used previously in Geocore. With our current search strategy, this test can only be performed on the short proteins, so we used 1PAO, 1EDP, 1TER, 1ANS, and 1DEP.

Acknowledgments

We appreciate the support of Ajinomoto Co. and support of a STAR biotechnology grant from the University of California.

References

Aho A, Hopcroft J, Ullman J. 1974. *The design and analysis of computer algorithms*. Reading, MA: Addison-Wesley.

Boczko EM, Brooks C. 1995. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science 269*:393–396.

Covell DG. 1992. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins 14*:409–420.

Covell DG. 1994. Lattice model simulations of polypeptide chain folding. *J Mol Biol 235*:1032–1043.

Dill KA. 1997. Additivity principles in biochemistry. *J Biol Chem 272*(2): 701–704.

Hinds D, Levitt M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol 243*:668–682.

Karplus PA. 1996. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci 5*:1406–1420.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding: I. lattice model and interaction scheme. *Proteins 18*:338–352.

Kuntz I, Crippen G, Kollman P, Kimelman D. 1976. Calculation of protein tertiary structure. *J Mol Biol 106*:983–994.

Levitt M, Warshel A. 1975. Computer simulation of protein folding. *Nature 253*:694–698.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 227*:876–888.

Monge A, Friesner R, 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.

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 22*:81–99.

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 18*:267–280.

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

Yue K, Dill KA. 1996. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci 5*:254–261.