Predicting Polypeptide and Protein Structures from Amino Acid Sequence: Antlion Method Applied to Melittin

TERESA HEAD-GORDON* and FRANK H. STILLINGER

AT&T Bell Laboratories, Murray Hill, New Jersey, 07974

SYNOPSIS

This report continues to explore the use of a strategy known as the antlion method for predicting polypeptide and protein structure. The method involves deformation of a biopolymer's potential energy hypersurface in order to retain only a single minimum, near to the native structure. The vexing multiple minimum problem thus is relieved, and the deformed hypersurface constitutes a key element in three-dimensional structure predictions with atomic resolution. In this more demanding pilot study, we provide evidence that the antlion method is capable of dramatically simplifying the surface of polypeptides by successfully predicting the native form of the naturally occurring 26-residue polypeptide melittin. The systematic hypersurface modifications employed in our previous work have been used again for this case, but have been supplemented by the output of a suitable neural network. This neural network involves a new feature: the use of amino acid biophysical scales for improving the secondary structure prediction accuracy of simple perceptrons.

INTRODUCTION

A central component of the protein-folding problem[1,2] is identification of the native state conformation. While the overall protein-folding problem encompasses understanding of the thermodynamic driving forces that act on the unfolded states as well as on the native protein,[3] and of the kinetic pathway by which the native state is obtained,[4-9] in its most streamlined version the task is to predict the full three-dimensional arrangement of the protein molecule, given only its primary structure (amino acid sequence) and the solvent conditions (composition, temperature, and pressure). Difficulties that must be faced stem from (a) the complexity of the proteins' intramolecular force field, (b) quantitative uncertainty about the nature of solvation for arbitrary conformation, and (c) the existence of many local minima in the solvent-averaged free energy hypersurface whose number apparently rises in roughly exponential manner with the number of amino acid residues. In spite of these difficulties, substantial effort has been devoted to resolution of the protein-folding problem, and this has produced a very large scientific literature devoted to the subject.[1-17]

In a recent manuscript[10] we began to explore a strategy, the "antlion method," that was devised specifically to relieve difficulty (c) above. It takes its name from a family of subterranean insects that lie in wait at the bottom of victim-entrapping basins. The ultimate objective of this method is to simplify the free energy (or potential energy) hypersurface for any polypeptide or protein so that only a *single* basin (and minimum) remains. Furthermore, the remaining minimum should occur close in configuration to that of the initial-hypersurface native-structure minimum. Optimization then proceeds in three stages: replace the complicated "real" hypersurface by its simplified variant, optimize on the modified hypersurface, and finally optimize on the real hypersurface, using the optimized structure found from the second stage as an initial guess, to locate its native structure minimum. Feasibility of this approach was supported by specific calculations on the blocked alanine dipeptide and the blocked alanine tetrapeptide.[10] Hypersurface modification for the former converted a 20-minimum topography

Biopolymers, Vol. 33, 293-303 (1993)

CCC 0006-3525/93/020293-11

* Present address: Lawrence Berkeley Laboratories, Berkeley, California 94720.

(or 40, counting mirror image structures) to the required single-minimum topography, while several hundred minima for the latter were collapsed to a single minimum as required.

The present paper is devoted to a small and simple, but nonetheless more demanding, test of the antlion method, specifically its capacity to predict with atomic resolution the native form of the naturally occurring 26-residue polypeptide melittin[18]: Gly-Ile-Gly-Ala-Val-Leu-Lys-Val-Leu-Thr-Thr-Gly-Leu-Pro-Ala-Leu-Ile-Ser-Trp-Ile-Lys-Arg-Lys-Arg-Gln-Gln. In this last respect our method stands in distinct contrast to lattice models[12,13] and to α-carbon representations.[17] The systematic hypersurface modifications employed in our previous paper[10] have been used again, but have been supplemented by the output of a suitable neural network. As reported in detail below, the prediction for melittin agrees satisfactorily with the experimental structure.[18] While melittin is quite simple structurally, it provides a pilot study that demonstrates the following points: (1) it describes the full implementation of the antlion strategy, where neural networks are used to guide the design of penalty functions; (2) it demonstrates the ability of the antlion method to overcome the multiple minimum problem (melittin has $\sim 10^{26}$ minima in the space of the backbone degrees of freedom alone), so that only the minimum near the native structure is retained; and (3) this new method demonstrates promise for future antlion method applications to more difficult tertiary structures.

The following section describes the generic potential energy model that we utilize as a test bed for the further development of the antlion method. The third section then introduces the antlion method, and reprises the elementary penalty functions developed earlier to modify the alanine dipeptide and tetrapeptide hypersurfaces,[10] and which we again use for melittin. The section after that presents our neural network formalism that is used to control the secondary structure penalty functions; this subsidiary role differs fundamentally from the direct predictive role usually assigned to neural networks in the protein folding problem. Our specific calculations for melittin appear in the fifth section. Conclusions and discussions reside in the final section.

POTENTIAL ENERGY FUNCTION

The empirical potential energy function used as the objective function Φ in this study has the form

$$\Phi = \sum_{i}^{\#\text{bonds}} k_{bi}(b_i - b_{i0})^2 + \sum_{i}^{\#\text{angles}} k_{\theta i}(\theta_i - \theta_{i0})^2 + \sum_{i}^{\#\text{improper}} k_{\tau i}(\tau_i - \tau_{i0})^2 + \sum_{i}^{\#\text{torsions}} k_{\omega i}[1 + \cos(n_i\omega_i + \delta_i)] + \sum_{i<j}^{N}\sum^{N} \times \{Cq_iq_j/r_{ij} + \epsilon_{ij}[(R_{ij}/r_{ij})^{12} - 2(R_{ij}/r_{ij})^6]\} \quad (1)$$

We have used the parameters of the extended atom representation (version 19) of CHARMM.[19] The first four terms refer to the chemical bond connectivity. The bond, bond angle, and improper torsion deformations are represented as harmonic potential functions with force constants k_b, k_θ, k_τ [the Hooke's law factor of $\frac{1}{2}$ has been factored into the force constants in Eq. (1)], and equilibrium values of b_0, θ_0, and τ_0, respectively. The torsional potential is represented as a Fourier cosine expansion, where k_ω is the force constant, δ is the phase, and n is a multiplicity factor that allows for inclusion of the higher harmonics. We note that in our application only one dihedral term is utilized for rotation around a given bond. The nonbonded terms in Eq. (1) are modeled as a sum of pairwise coulomb electrostatic and Lennard–Jones hard interactions. The Lennard–Jones cross-interaction parameters are evaluated using conventional simple mixing rules[20]:

$$\epsilon_{ij} = [\epsilon_{ii} \times \epsilon_{jj}]^{1/2}$$
$$R_{ij} = (R_{ii} + R_{jj})/2 \quad (2)$$

In addition, the electrostatic interactions are scaled by a factor $C = 0.4$ when the pair under consideration is separated by three bonds. A cutoff of 7.5 Å is used for the evaluation of all pair interactions, using a shifting function[19] to smooth the energy and derivatives. For further details of the specific CHARMM parameters, see Ref. 19.

One aspect of a complete solution to the protein-folding problem involves the quantitatively accurate description of the free energy hypersurface of the solvated biopolymer. As we have indicated in the Introduction, we have chosen not to address this issue at this time, since the antlion strategy is directly transferrable to more quantitative free energy (or potential energy) functions as they become available. However, we feel compelled to delineate the reasons why the native structure minimum that we isolate on the modified potential energy hypersurface, and that we ultimately converge to on the

empirical unmodified potential energy surface,[10,19] should plausibly resemble that of the in vivo structure.

We begin by noting that our calculations in effect are done in the gas phase; no attempt was made to include an obvious solvent component, such as a dielectric constant of 80, an r-dependent dielectric behavior, or explicit configurations of molecular water. The adequacy of the nominally gas phase potentials themselves as structural predictors of the native structure deserves comment. Recent studies indicate that most empirical potential energy functions[19,21-23] show reasonable structural agreement, although poor relative energy ordering of the minima,[19,23] when compared with high level ab initio calculations[24,25] for the gas phase ϕ,ψ surfaces of the hydrogen-blocked glycine and alanine dipeptides. The differences observed between the ab initio results[24,25] and empirical potential functions[19,23] may be due to the fact that the latter have been parameterized to reproduce the structural and energetic aspects of x-ray experimental data. While crystalline forces might have been thought to distort the structure from that corresponding to its structure in solution, preliminary nmr structural studies indicate that the crystal structure is a good approximation to the solution structure[26,27] in those cases where a comparison could be made. While we remain sensitive to the lack of a solvent component in our present choice of potential energy function, we believe the empirical force fields provide an adequate, although far from perfect, representation of the native protein structure in vivo.

THE ANTLION STRATEGY

The antlion strategy involves the deformation of the objective function hypersurface Φ in Eq. (1) in such a way that a preselected minimum (which is designed to be a close approximation to the native structure minimum) forms the dominant basin on the surface. Thus starting at any initial configuration of a biopolymer, for example the fully extended conformer (all ϕ,ψ pairs defined as 180°, −180°), any minimization technique will converge to this single remaining minimum. Once this relevant area of configuration space has been reached, regeneration of the original surface is achieved by using the unmodified (objective) function Φ to refine the structure.

The modification of the objective function is accomplished by the addition of penalty functions. In the case of alanine dipeptide and tetrapeptide, we have found three useful types of penalty functions.[10] In most cases, we desire the elimination of all minima where particular amino acids have the wrong chirality, i.e., the D configuration. We have used the following elementary penalty function to bias in favor of the L configuration:

$$V_\tau = k_\tau(\tau - \tau_0)^2 \qquad (3)$$

where τ corresponds to the torsions C_α-N-C-H_α and C_α-N-C-C_β, and τ_0 is appropriate for L isomers. In addition, the elimination of all minima where peptide groups are in the *cis* conformation is generally desirable. We note that the peptide torsion potential usually used

$$V_p = k_p[1 + \cos(2\omega + \pi)] \qquad (4)$$

possesses minima at both $\omega = 0$ and π. The obvious modification of Eq. (4) to favor the *trans* form is to change the multiplicity factor of 2 to 1, and to change the phase from π to 0. A similar modification is easily implemented for the retention of *cis* peptides if so desired. In order to maintain the original curvature at the minimum, we use a force constant of $4k_p$ in the modified version of Eq. (4).

The knowledge that an amino acid is in a particular type of secondary structure allows the construction of penalty functions using the definition of that secondary structure. For example, an amino acid i, which is α-helical in a particular polypeptide or protein, would ideally require the formation of a hydrogen bond between residue i and $i + 4$ of 1.9 Å, and the adoption of backbone dihedral angles ϕ and ψ of −57° and −47°, respectively. Similar ideas can be extended to other types of secondary structure such as reverse turns and β-sheets.

We have demonstrated[10] that the following penalty function

$$V_{\phi\psi} = k_\phi[1 - \cos(\phi - \phi_0)] + k_\psi[1 - \cos(\psi - \psi_0)] \qquad (5)$$

successfully restrains the backbone dihedral angles to any desired ϕ_0,ψ_0, with appropriately chosen k_ϕ and k_ψ. The addition of this set of penalty functions allowed us to maintain one and only one minimum on the alanine dipeptide and tetrapeptide ϕ,ψ surfaces for all ϕ_0 and ψ_0 pairs of interest.[10]

In addition to the backbone dihedral angle restraints, we also utilize intramolecular hydrogen-bond penalty functions for the formation of secondary structures such as helices, turns, and sheets. We

will demonstrate in the fifth section that an electrostatic "reward" function

$$V_{2^0} = q_i q_j / r_{ij} \tag{6}$$

provides a useful modification of the original objective function [Eq. (1)] of melittin, so that hydrogen bonds appropriate to an α-helix are retained.

The alanine dipeptide and tetrapeptide examples seem to imply that prior knowledge of the secondary and tertiary structure of globular proteins is required in order to implement the antlion approach for these larger biopolymers. It would hardly be a useful tertiary structure predictor if this were the case. In order to avoid such circularity, we therefore adapt our antlion strategy to use neural networks as a guide for designing penalty function parameters that retain only the native globular protein minimum. We wish to emphasize the distinction between our use of neural networks, and that conventionally required of neural networks in the protein-folding area.[17,28–31] For the latter, the outputs of the network are the direct structure predictions, whether they be secondary structure predictions[28–31] or residue contact distance classification.[17] In our approach, neural networks serve as an intermediary between the amino acid sequence and structure prediction, since they are intended to be used as a predictor for the penalty parameters only. Minimization first on the modified potential hypersurface and then on the unmodified hypersurface serves as the tertiary predictor. Local violations of the neural network predictions then become possible as the entire system seeks and finds its optimal final structure. In this respect our approach accommodates the presence of locally frustrated interactions in the interests of attaining a global minimum tertiary structure.

NEURAL NETWORK DESCRIPTION

Neural network algorithms for performing learning tasks such as pattern recognition are conceptually based on the structure and function of the central nervous system.[32] In the context of the protein-folding problem, neural network algorithms are required to predict patterns of secondary and tertiary structure of the native protein (neuronal response, or output) from the amino acid sequence (sensory input to the network).

The topology of the neural network we have used to predict the backbone dihedral penalty functions for melittin is that of the simple perceptron, also known as feed forward-back propagation networks with no hidden layers.[32] In this case, each amino acid of a protein sequence is represented by a small set of input neurons that is directly connected, or fed into, output neuron(s) representing a secondary structure classification. The small set of input neurons generally correspond to the amino acid whose most likely secondary structure is being predicted, while the remainder supply a context (or window) of n amino acids (8 in our study) preceding and succeeding this amino acid along the backbone. The learning, or training, phase of the neural network algorithm involves minimizing the function

$$E = \sum_i^N \sum_j^M (O_{0j}^i - O_{cj}^i)^2 \tag{7}$$

where M is the number of output units, N is the number of presented input patterns, O_0 is the observed secondary structure output, O_c is the calculated output. The calculated output is determined as follows:

$$A_{cj}^i = \sum_k^L w_{jk} I_k^i + b_j \tag{8}$$

and

$$O_{cj}^i = 1/[1 + \exp(A_{cj}^i)] \tag{9}$$

where L is the number of input units, I is the input, w_{jk} is the weight of the connection between the input neuron k and output neuron j, and b_j is the bias associated with the output neuron j. We use a steepest descent algorithm for minimizing the function in Eq. (7) with respect to the free parameters w_{jk} and b_j. The parameters w_{jk} and b_j are updated (or "back propagated" through the network from output to input) by the following derivative expression:

$$\Delta w_{jk} = -\gamma \partial E / \partial w_{jk} \tag{10}$$

$$\Delta b_j = -\gamma \partial E / \partial b_j \tag{11}$$

where γ is a damping or "learning" factor,[32] taken to be 0.0002 in this study.

We have tried to exploit physically motivated ideas concerning input and output representations, in order to improve the secondary structure prediction accuracy of our neural networks. Input and output representation involves encoding biophysical properties into the amino acid sequence (input) and secondary structure (output). For example, each of

the 20 amino acids could be represented by a 5-bit binary number ranging from 00001 to 11111. To reflect a physically relevant property, such as hydrophobicity for example, the amino acids would be assigned a 5-bit number depending on where the residue sits in the hydrophobic scale.[33] Isoleucine being least hydrophilic would be assigned the 5-bit number 00001, while the most hydrophilic amino acid arginine would be assigned the 5-bit representation 10111. The "blanks" in the window at polypeptide chain ends might be given a value of 11111, with the idea that chain ends are charged and solvent exposed, and therefore most hydrophilic. Similarly, output assignments could be ordered to reflect hydrogen-bond formation local in sequence (helices and turns), nonlocal in sequence (ladders, sheets), and no hydrogen-bond formation (bends and coil). The preliminary results we provide in this work indicate that these ideas of biophysical representation have noteworthy impact on network predictions of secondary structure.

As a straightforward implementation of this general idea, we have designed the following highly simplified network. The input representation for each amino acid is a 5-bit binary number ordered to reflect one of the following scales: an α-helix promotion ordering of the amino acids deduced from substitutions of the commonly occurring residues into a coiled coil,[34] an α-helix promotion scale based on a statistical analysis of 60 proteins,[35] and a random scale generated from a normal distribution. The three scales are presented in Table I. The output is designed to be one neuron that is "helical" when on (output value of 1), and "nonhelical" when off (output value of 0). The choice of a helix/no helix network is motivated by two points: first, melittin is largely α-helical, and second, it provides a simple test of the relevance of the biophysical scale representation in our neural network.

The very simple network described above (context of 17, 5 bit input, no hidden layers, 1 bit output) was trained on a subset[29] of the data base and secondary structure identifications of Kabsch and Sander[36,37]; we have not in any way exploited homologies, criteria for acceptable refinement of the x-ray data, etc. We also note that the Kabsch and Sander secondary structure identifications are only objective to the extent of their definition of secondary structure—those that are fully hydrogen bonded. Different conclusions about the presence of secondary structure types, or their absence, for a particular amino acid in a data-base protein may be reached by different criteria. The Kabsch and Sander data base serves the immediate purpose of providing self-consistent results in the neural network learning process for the study presented here. There is certainly merit for critically assessing the deficiencies of training data bases in the future, since this will contribute to the accuracy of the final predicted polypeptide or protein structure. We optimize the decision of whether the neuron is on or off on the training set (after the weights and biases have been optimized), by defining a threshold t, which gives a maximum in the correlation coefficient, or predictive confidence. We have used the following correlation coefficient definition[29,31]:

Table I Input Representation: α-Helix Promotion

Residue	Levitt	O'Neill and DeGrado	Random
Met	00001	00101	00001
Glu	00010	01010	10000
Leu	00011	00100	01101
Ala	00100	00001	01100
Gln	00101	01001	00010
Lys	00110	00011	00100
His	00111	10010	01110
Cys	01000	01011	00110
Phe	01001	00111	01001
Asp	01010	01110	01000
Trp	01011	00110	10001
Ile	01100	01100	10010
Arg	01101	00010	00101
Val	01110	01111	01111
Asn	01111	10001	01010
Ser	10000	01000	00011
Thr	10001	10000	00111
Tyr	10010	01101	01011
Gly	10011	10011	10011
Pro	10100	10100	10100

$$C_\alpha = \frac{p_\alpha n_\alpha - u_\alpha o_\alpha}{[(n_\alpha + u_\alpha)(n_\alpha + o_\alpha)(p_\alpha + u_\alpha)(p_\alpha + o_\alpha)]^{1/2}} \tag{11}$$

where p_α is the number of α-helical output patterns predicted correctly, n_α is the number of nonhelical outputs rejected correctly, u_α is the number of underpredicted helical output patterns, and o_α is the overprediction of helical patterns. The optimized network of weights, biases, and threshold is then presented with the testing data base[29] (the remaining proteins of the Kabsch and Sander data base[36,37] not present in the training set). The predictive capacity, defined by the percentage of helix predicted

correctly, percentage of nonhelix predicted correctly, and the correlation coefficients for the training and testing sets of each of the three types of networks, are listed in Table II. As is evident from these results, the random scale representation performs significantly more poorly than the two scales encoding the biophysical property of helix promotion. Furthermore, the O'Neill and DeGrado scale[34] does somewhat more poorly than the Levitt scale[35]; this is likely due to the fact that the O'Neill and DeGrado scale[34] is inferred from a more artificial environment of a guest amino acid site in a short polypeptide coiled coil, which is highly idealized in comparison to the data base of globular proteins used in the training and testing set. For the case of melittin, the random scale predicts only 3 residues to be helical out of the 22 possible helical residues, while the DeGrado scale predicts 6 and the Levitt scale 11. Interestingly, neither the O'Neill and DeGrado or Levitt scale predict the first half of the amino acid sequence of melittin to be helical; however, this is due to the fact that the magnitude of the output for this half of the sequence falls just below the hard thresholds optimized for these two scales.

It is appropriate to reemphasize at this point that in our application the neural networks themselves do not serve as the predictor, but merely provide a means for defining appropriate penalty function parameters. In the following section we show that the deficiencies of the neural network (i.e., only 11 of the 22 helical amino acids of melittin predicted correctly with our best network) need not preclude successful secondary or tertiary prediction. We provide a demonstration of this point by folding the small polypeptide melittin in the next section.

RESULTS FOR MELITTIN

Figure 1 displays a minimized, polar hydrogen, molecular mechanics structure, which represents the observed native state of melittin. The heavy atoms of the 2.0 Å crystal structure[18] were provided with polar hydrogens (i.e., extended atom representation for methyl, methine, etc.) so that excluded volume and geometric considerations are satisfied. The resulting hydrogenated structure was minimized with large harmonic constraints on the heavy atoms for several hundred steps using adapted basis Newton Raphson (ABNR) with the molecular mechanics package CHARMM.[19] The constraints were iteratively reduced by 20% of their initial value, and the structure minimized for several hundred steps at each constraint value, until no constraints remained. The resulting rms comparison of the heavy atom crystal structure and the heavy atom minimized structure is given in Table III. The melittin crystal structure[18] can be described as helical[36] for residues 2–10 and 13–25, with a turn or bend at residues 11 and 12, while the first and last residue reside in a random coil geometry. The minimized hydrogenated structure exhibits helical segments for amino acids 2–10 and 13–21. This structure possesses a classic type III turn at residues 11 and 12. Residues 22–26 are classified as random coil geometries, i.e., all 5 residues show ϕ,ψ values far removed from the α-helix conformer, and only 3 residues are involved in 2 hydrogen bonds in this region (19–23 and 22–26), compared to the 5 residues involved in 5 hydrogen bonds in the original crystal structure (18–22, 19–23, 20–24, 21–25, 22–26).

The starting structure for our antlion procedure is the minimum closest to the fully extended form of melittin. We define the fully extended structure to have idealized geometries for chemical bonds and angles, and all dihedrals to be in their optimal rotamer minimum (for example, ϕ,ψ backbone values of $-180°,180°$, respectively). This idealized structure has many bad nonbonded contacts, and hence is relaxed using ABNR to a nearby minimum defined by a converged gradient of 0.005 kcal/(mole·Å). This relaxed structure, presented in Figure 2, is the input for our antlion procedure. Notice that it differs drastically from the native structure shown in Figure 1.

The antlion strategy for modifying the potential energy surface of melittin is as follows. The ϕ_0,ψ_0 penalty parameters [Eq. (5)] were assumed to be

Table II Secondary Structure Prediction on Kabsch and Sander Data Base

	Levitt	O'Neill and DeGrado	Random
Structural probe	Train/test	Train/test	Train/test
% α correct	69/64	54/58	63/53
% else correct	70/67	78/74	68/65
C_α	0.36/0.29	0.31/0.30	0.28/0.17

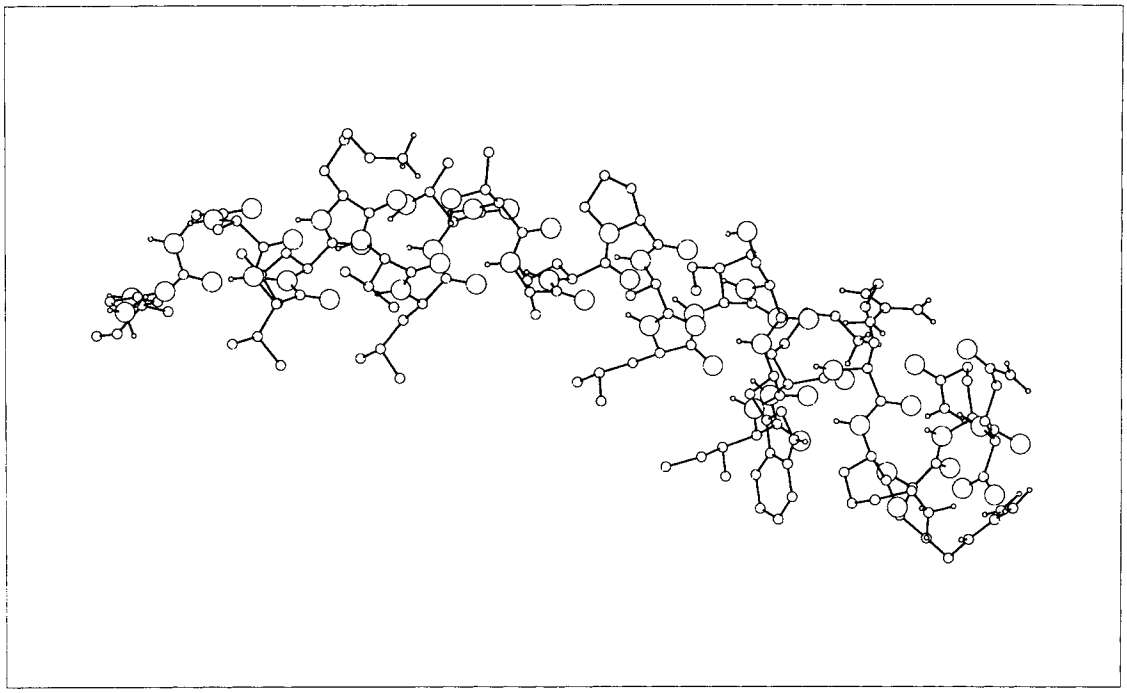

Figure 1. Mellitin native structure. The constrained minimized structure of the x-ray diffraction structure by Terwilliger and Eisenberg. The structure is characterized by helical conformations for residues 2–10 and 13–21; there is a type III turn at amino acids 11 and 12. The remaining residues are of a coil configuration, although some secondary structure is present.

−57°,−47°, which is the ideal α-helix backbone conformation.[38] The force constants k_ϕ and k_ψ are set equal to the output from the network discussed above using the Levitt scale, which is a real number between 0.0 and 1.0, and then scaled by a factor of 100 kcal/mole. This gives force constants that fall between the magnitude of the peptide torsions and bond angle force constants appearing in Eq. (1) (8–55 kcal/mole). We also invoke the formation of hydrogen bonds between the backbone oxygen of residue i and the backbone hydrogen of residue $i + 4$ by the use of Eq. (6), where $q_i = -q_{i+4}$ is the direct network output ($0.08e^-$ to $0.55e^-$); all side-chain atom charges were set to 0.0. In addition, we have included the penalty functions corresponding to the elimination of D-isomers [Eq. (3)] and *cis* peptide [Eq. (4)] minima in this calculation for melittin, although these functions are minimized based on our extended structure starting guess.

The minimized structure on the modified surface was then used as the starting structure on the unmodified surface, and minimized to the same tol-

Table III RMS Difference Between Experimental and Antlion Structures

Residues	Crystal/Minim RMS (Å)	Crystal/Antlion RMS (Å)	Minim/Antlion RMS (Å)
1–26	2.217	2.457	2.535
Backbone, 1–26	1.511	1.220	1.963
α-Carbons, 1–26	1.540	1.282	1.965
2–25	2.163	2.340	2.311
Backbone, 2–25	1.403	1.208	1.772
α-Carbons, 2–25	1.442	1.259	1.753

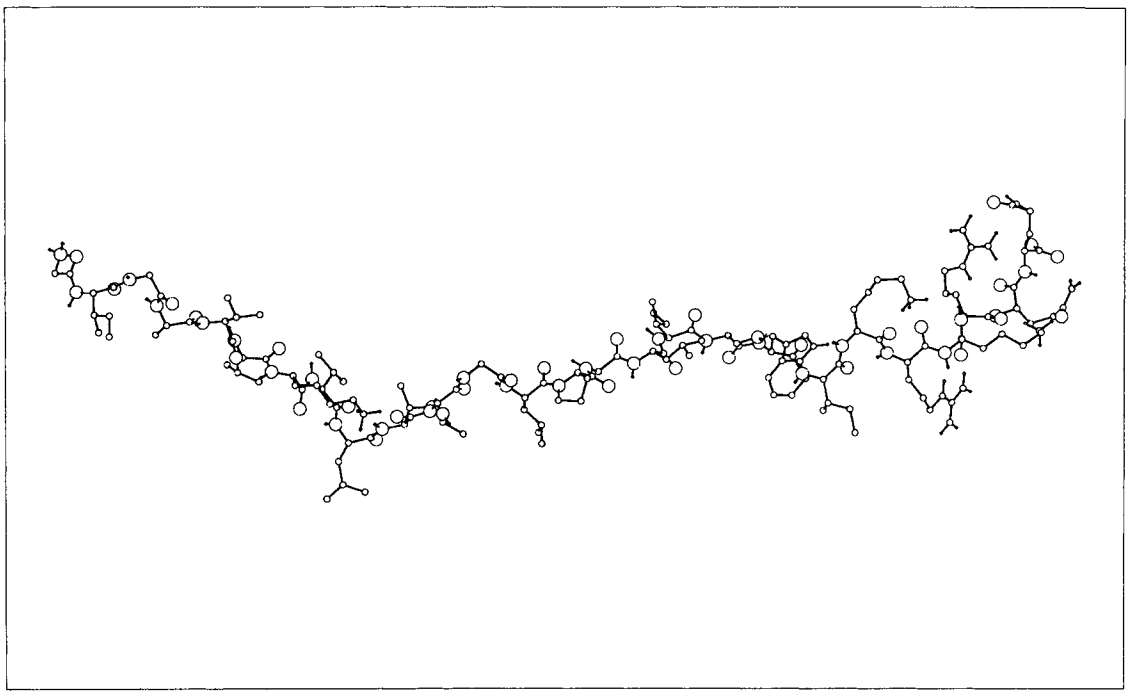

Figure 2. Melittin extended, minimized structure. This structure was used as an initial guess for the antlion procedure. There is no obvious secondary or tertiary structure present in this structure.

erance of 0.005 kcal/mole · Å. The resulting minimum on the unmodified surface, depicted in Figure 3, should be contrasted to the relaxed extended conformer of melittin in Figure 2. Clearly the antlion method has eliminated the extended conformer local minimum. The antlion folded structure shows helical segments for amino acids 2–10 and 13–21. Residues 11 and 12 are in a bend conformation, while the remaining nonhelical residues 23–26 exhibit a coil configuration. Residue 22 has ϕ,ψ values in the α-helical region, and is hydrogen bonded to residue 26 (although not to 18). Residues 23–26 exhibit nonhelical ϕ,ψ values and hydrogen bonds between residues 19–23, 20–24, and 22–26.

A comparison of the crystal structure with the folded structure of melittin obtained from the antlion procedure is shown in Figure 4; there is remarkable similarity for the backbone conformation. The rms differences between our folded structure and the crystal structures (heavy atoms and hydrogenated) are given in Table III, with our best value being 1.21 Å for a comparison of the backbone atoms of residues 2–25 (i.e., excluding the coiled ends). The rms difference between the entire antlion structure with the crystal structure, 2.54 Å, is close to the resolution of the experiment, 2.0 Å.[18]

There are four important points to be made at this juncture. The first is that the neural network outcome itself would only have predicted that 11 out of the possible 22 helix residues are helical. A simple scaling of the output as a penalty function improves this prediction so that 19 out of 22 are helical, due to the fact that 5 of the 7 predicted directly by the network to be nonhelical, sat marginally below the threshold.

Second, while there is some sensitivity of the quality of the predicted structure to the magnitude of the penalty function scale factor, there are well-defined reasons for choosing the scale factor of 100 kcal/mole. We have found that the largest barrier to eliminate in the smoothing process is that due to bond angle strain; thus penalty function force constants must be the same magnitude in order to compete with these barriers. For example, we have found that the rms deviation of the predicted structure degrades when the output is scaled by 50 kcal/mole, which is due to penalty function force constants which are too soft (4–27 kcal/mole) to compete with the bond angle potential. The use of the O'Neill and DeGrado and random networks, scaled to give force constants in the bond angle range, do not predict the melittin structure as well as the Levitt scale (rms

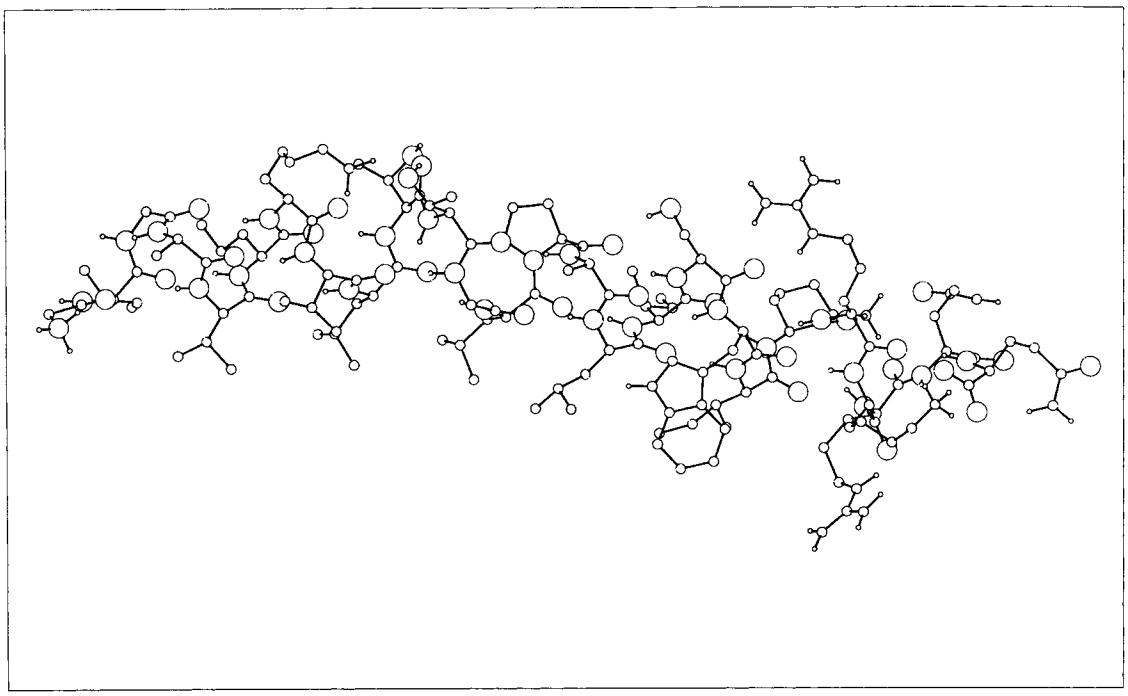

Figure 3. Melittin folded structure obtained from the antlion procedure. The antlion folded structure is characterized by helical segments 2–10 and 13–21, with a bend at residues 11 and 12. The remaining residues, 22–26, are classified as random coil, although some secondary structure is present (see text for details).

deviations of 1.68 and 2.13 in the backbone degrees of freedom, respectively, and 2.92 and 3.20 for all degrees of freedom, respectively).

Third, the use of the ideal ϕ_0,ψ_0 values of $-57°,-47°$ and hydrogen bonds between residues i and $i+4$ seems to assume the correct structure, and not predict it; however, the antlion method successfully finds the end residues 1 and 26 to be far removed from the helix conformation, and defines an appropriate turn or bend at residues 11 and 12.

Last, a comparison of side-chain conformations between the crystal structure and the antlion folded structure (Figure 4) clearly indicates that the native structure minimum and that found by the antlion procedure may not be the same. In fact, there are multiple minima on the modified hypersurface in the space of the sidechain degrees of freedom. However, as we have already discussed, the modified surface is believed to retain only a very small subset of the original number of minima in the subspace of the backbone conformations. We are not overly concerned with the multiple minimum problem in the space of side-chain conformations since a good prediction of the backbone limits the conformational possibilities for the side chains, thereby allowing exhaustive searches in this subspace.[39] It is also conceivable that other neural network schemes could be devised for the side-chain degrees of freedom.

DISCUSSION AND CONCLUSIONS

In summary, we have implemented a strategy known as the antlion method for greatly simplifying polypeptide and protein potential energy hypersurfaces in order to retain only one conformationally distinct minimum corresponding to the native structure. In this work, we have adapted the antlion strategy to incorporate neural networks, and have demonstrated this adaptation for successfully predicting the structure of the 26-residue polypeptide, melittin. We emphasize again that the output of the neural networks themselves are not used as the structure predictor; instead they serve the purpose of guiding the selection of penalty functions that deform the objective function hypersurface to retain only that minimum corresponding to the native structure. In addition, we have also shown that the use of biophysical scales in the design of neural networks for secondary, and possibly tertiary, structure prediction

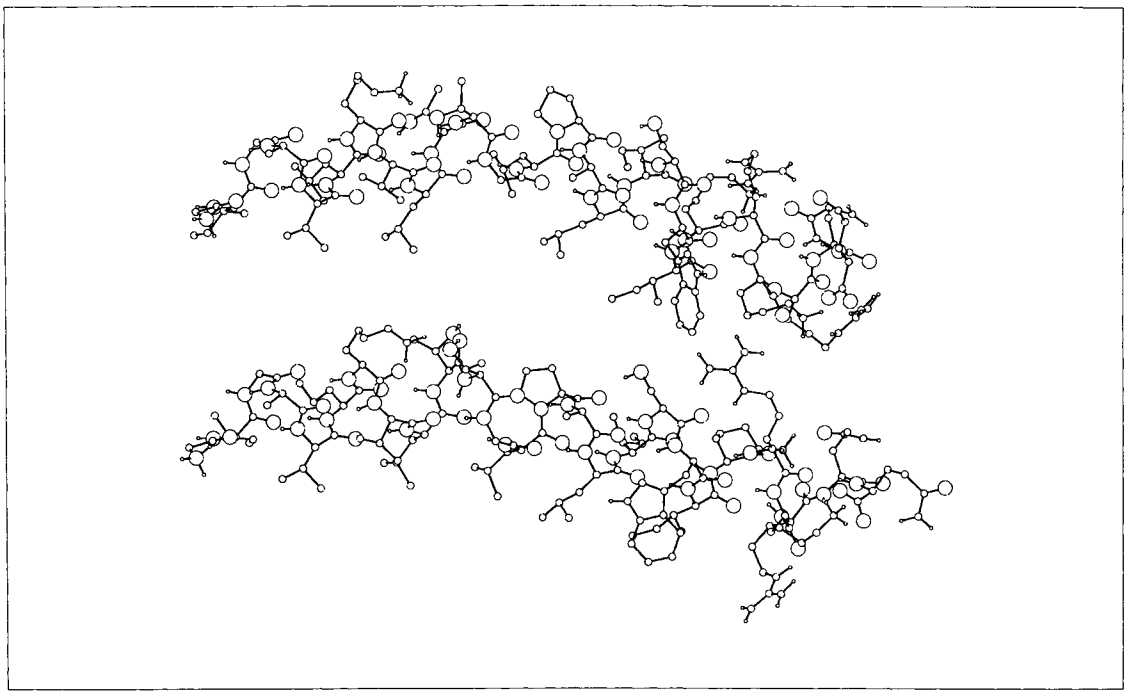

Figure 4. A comparison of the minimized crystal structure and the antlion folded structure. An overall rms difference of 2.45 Å between these structures is observed, while the backbone degrees of freedom show an rms difference of 1.2 Å.

may provide some useful improvements over those network designs currently used in the protein folding area.

While our previous paper[10] has indicated that the antlion approach is feasible for di- and tetrapeptides, the current study has shown that the method can be successfully applied to significantly larger polypeptides and proteins, as exemplified by the small toxin protein melittin, where a brute force search procedure becomes intractable. It should be emphasized that although the case of melittin is a significant step forward, in no way do we claim complete solution to the problem of protein structure prediction. First, little tertiary structure is present in the case of melittin, so that success was relatively easily attainable. Second, other predictive strategies will be needed to supplement the very simple helix neural network algorithm presented here, in order to move onto proteins with much richer tertiary structure than that of melittin. We currently are investigating other biophysical scales for the improved prediction of β-sheets and β-turns, in addition to α-helix prediction. We are additionally pursuing the use of Hopfield-like neural networks[40] for the prediction of hydrogen-bond and/or disulfide-bond matrices. We also believe it is possible to improve the data base by exploiting homologies between the training and testing sets.

Once these algorithmic components are in place, we foresee the following flow diagram for the antlion approach for predicting tertiary structure in any protein:

1. amino acid sequence
   ↓ Neural Networks
2. 2° and/or 3° structure penalty parameters
   ↓ Define modified surface
3. Minimization on modified surface using extended conformer as starting structure
   ↓ Regenerate original objective function
4. Minimization on unmodified surface using the minimized structure found from point 3 as the starting structure
   ↓ Converge structure to strict tolerance
5. Predicted structure determined with atomic resolution

Thus, the most ambitious scenario is a method, which for any polypeptide or protein, predicts atomic resolution structures using the amino acid sequence as sole input.

We thank Dr. Lynn Jelinski for many useful interactions. We also thank Professor Charles Brooks III for use of the program CHARMM. Finally, we thank Dr. Peter Mirau for his help in providing the figures in this paper.

REFERENCES

1. Gierasch, L. M. & King, J., eds., (1990) *Protein Folding: Deciphering the Second Half of the Genetic Code.* American Association for the Advancement of Science. Washington, D.C.
2. King, J. (1989) *Chem. Eng. News* **67,** 32-54.
3. Chan, H. S. & Dill, K. A. (1991) *Ann. Rev. Biophys. Biophys. Chem.* **20,** 447-490.
4. Baum, J., Dobson, C. M., Evans, P. A. & Hanley, C. (1989) *Biochemistry* **28,** 7-13.
5. Creighton, T. E. (1977) *J. Mol. Biol.* **113,** 295-312.
6. Weissman, J. S. & Kim, P. S. (1991) *Science* **253,** 1386-1393.
7. Kuwajima, K. (1977) *J. Mol. Biol.* **114,** 241-258.
8. Roder, H., Elöve, G. A. & Englander, W. S. (1988) *Nature* **335,** 700-704.
9. Udgaonkar, J. B. & Baldwin, R. L. (1988) *Nature* **335,** 664-669.
10. Head-Gordon, T., Stillinger, F. H. & Arrecis, J. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 11076-11080.
11. Gō, N. (1983) *Ann. Rev. Biophys. Bioeng.* **12,** 183-210.
12. Kolinski, A., Skolnick, J. & Yaris, R. (1988) *Proc. Natl. Acad. Sci. USA* **83,** 7267-7271.
13. Lau, K. F. & Dill, K. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6388-6392.
14. Levitt, M. (1976) *J. Mol. Biol.* **104,** 59-107.
15. Piela, L. & Scheraga, H. A. (1988) *Biopolymers* **26,** S33-S58.
16. Shakhnovich, E. I. & Gutin, A. M. (1989) *Biophys. Chem.* **34,** 187-199.
17. Friedrichs, M. S., Goldstein, R. A. & Wolynes, P. G. (1991) *J. Mol. Biol.* **222,** 1013-1034.
18. Terwilliger, T. C. & Eisenberg, D. (1982) *J. Biol. Chem.* **257,** 6016-6022.
19. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4,** 187-217.
20. Kohler, F., Fischer, J. & Wilhelm, E. (1982) *J. Mol. Struct.* **84,** 245-250.
21. Jorgensen, W. L. & Tirado-Rives, J. (1988) *J. Am. Chem. Soc.* **110,** 1657-1666.
22. Momany, F. A., Klimkowski, V. J. & Schäfer, L. (1990) *J. Comp. Chem.* **11,** 654-662.
23. Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986) *J. Am. Chem. Soc.* **106,** 230-252.
24. Head-Gordon, T., Head-Gordon, M., Frisch, M. J., Brooks, C. L. & Pople, J. A. (1989) *Int. J. Quant. Chem. Quant. Biol. Symp.* **16,** 311-322.
25. Head-Gordon, T., Head-Gordon, M., Frisch, M. J., Brooks, C. L. & Pople, J. A. (1991) *J. Am. Chem. Soc.* **113,** 5989-5997.
26. Kline, A. D., Braun, W. & Wuthrich, K. (1986) *J. Mol. Biol.* **189,** 377-382.
27. Pflugarth, J. W., Weingard, G. & Huber, R. (1986) *J. Mol. Biol.* **189,** 383-386.
28. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Norskov, L., Olsen, O. H. & Petersen, S. B. (1990) *FEBS Lett.* **261,** 43-46.
29. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152-156.
30. Kneller, D. G., Cohen, F. E. & Langridge, R. (1990) *J. Mol. Biol.* **214,** 171-182.
31. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202,** 865-884.
32. Müller, B. & Reinhardt, J. (1990) *Neural Networks: An Introduction,* Springer-Verlag, Berlin-Heidelberg.
33. Clothia, C. (1976) *J. Mol. Biol.* **105,** 1-14.
34. O'Neill, K. T. & DeGrado, W. F. (1990) *Science* **250,** 646-651.
35. Levitt, M. (1978) *Biochemistry* **17,** 4277-4285.
36. Kabsch, W. & Sander, C. (1983) *FEBS Lett.* **155,** 179-182.
37. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.
38. Creighton, T. E. (1984) *Proteins, Structures and Molecular Properties,* W. H. Freeman, New York.
39. Lee, C. & Subbiah, S. (1991) *J. Mol. Biol.* **217,** 373-388.
40. Hopfield, J. & Tank, D. W. (1985) *Biol. Cybern.* **52,** 141.

Received December 18, 1991
Accepted April 20, 1992