Collective aspects of protein folding illustrated by a toy model

Frank H. Stillinger
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

Teresa Head-Gordon
Life Sciences Division, Lawrence Berkeley Laboratory, University of California, Berkeley, California 94720

(Received 14 April 1995)

A simple toy model for polypeptides serves as a testbed to illuminate some nonlocal, or collective, aspects of protein folding phenomena. The model is two dimensional and has only two amino acids, but involves a continuous range of backbone bend angles. Global potential energy minima and their folding structures have been determined for leading members of two special and contrasting polypeptide sequences, center doped and Fibonacci, named descriptively for their primary structures. The results display the presence of spontaneous symmetry breaking, elastic strain, and substantial conformational variation for specific embedded amino acid strings. We conclude that collective variables generated by the primary amino acid structure may be required for fully effective protein folding predictors, including those based on neural networks.

PACS number(s): 87.10.+e, 87.15.By, 82.90.+j

I. INTRODUCTION

The general protein folding problem continues to present formidable scientific challenges. In an effort to isolate and clarify a few aspects of protein folding phenomena, the theoretical community has introduced and examined several highly simplified, but still nontrivial, models. These have included a large family of lattice models [1–8], models based on spin glass concepts [9–12], and continuum polymer models that incorporate simplified backbone potentials and interresidue interactions [13–15]. The present paper concerns an example of the last of these three categories; it was recently created to explore the adaptation of neural networks to prediction of folding patterns of proteins from their primary structure (amino acid sequence along the backbone) [16,17].

While the immediate neighbors along the backbone of any amino acid residue exert considerable statistical influence on the conformation of that residue in the properly folded state, it is clear that other "nonlocal" effects are also significant. Collective features of the folding bring residues into contact that may be widely separated in the primary structure, i.e., along the backbone. The inability, in general, to anticipate these larger-scale aspects of folded structures has had a limiting effect on the predictive ability specifically of various neural network schemes that have been proposed [18–22].

The objective in the present work has been to illustrate the source and the complex influence of nonlocal, or collective, effects in protein folding using our very simple "toy model" [16]. Specifically, we have obtained the lowest-energy folded states for the leading members of two special and contrasting sequences of toy polypeptides. These include examples containing up to 55 residues. For reasons that will become clear, these sequences are named "center-doped" and "Fibonacci" sequences. It is our hope that results and interpretive comments offered below will have a stimulating influence in development of new theoretical tools to cope with the general protein folding problem.

Section II briefly defines the toy model and lists a few of its elementary properties. Sections III and IV present the global energy minima for the first few members respectively of the center-doped and Fibonacci sequences. Section V summarizes our conclusions based on those minima. Section VI contains some additional discussion and perspective on possible future research directions.

II. TOY MODEL

The calculations to follow are based on a two-residue (A,B), two-dimensional, linear polymer model [16]. Links between successive residues along the backbone have fixed length unity, but the backbone can bend continuously between any pair of successive links. The potential energy function Φ contains two kinds of contributions and for an n-residue molecule is written

$$\Phi=\sum_{i=2}^{n-1} V_1(\theta_i)+\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} V_2(r_{ij},\xi_i,\xi_j)\ . \tag{2.1}$$

Here $-\pi\le\theta_i<\pi$ is the backbone bend angle (away from linear) at nonterminal residue i, r_{ij} is the distance between residues i and j, and the discrete variables ξ_i denote residue species

$$\xi_i(A)=+1,\quad \xi_i(B)=-1\ . \tag{2.2}$$

The backbone bend potential V_1 has a simple trigonometric form (we use reduced units)

$$V_1(\theta_i)=\tfrac{1}{4}(1-\cos\theta_i)\ , \tag{2.3}$$

which has extrema at $\theta_i=0,\pm\pi$. The residue pair interactions V_2 (which only operate between unlinked resi-

dues) possess a species-dependent Lennard-Jones form

$$V_2(r_{ij},\xi_i,\xi_j)=4[r_{ij}^{-12}-C(\xi_i,\xi_j)r_{ij}^{-6}] ,$$
$$C(\xi_i,\xi_j)=\tfrac{1}{8}(1+\xi_i+\xi_j+5\xi_i\xi_j) . \quad (2.4)$$

The coefficient C is $+1$, $+\frac{1}{2}$, and $-\frac{1}{2}$, respectively, for AA, BB, and AB pairs, thus producing an intramolecular mix of strong attraction, weak attraction, and weak repulsion, roughly analogous to the situation in real proteins.

The presence of the backbone bend potentials V_1 in Φ causes the strictly linear conformation (all $\theta_i=0$) to be a local Φ minimum for all n and for all residue sequences. But if $n>6$, regardless of the residue sequence, the global minimum corresponds to a nonlinear folded geometry that takes advantage of attractive AA and BB pair interactions, while avoiding a substantial penalty in repulsive AB pair interactions. The delicate interplay between competing contributions to Φ creates great diversity in the structures of optimally folded toy polypeptides and in particular causes some, but not all, point mutations to induce substantial shifts in the most stable folding pattern [16,17].

One last element of model simplification should be noted. Unlike real proteins, the backbone of the toy model molecules has no preferred directionality: forward and back are basically indistinguishable. This has a minor effect on the enumeration of fundamentally distinct toy model proteins containing n residues. If $n\equiv 2m$ is even, the number of distinct molecules is $2^{m-1}(2^m+1)$; for odd $n\equiv 2m+1$ the number is $2^m(2^m+1)$. If the backbone were directed, of course, the result would always be 2^n. Several extensions of the toy model are possible (such as dipoles embedded in the A's and B's), which would remove this bidirectionality.

III. CENTER-DOPED SEQUENCE

Not only does the number of distinct molecules rise essentially exponentially with n, but so too does the computational difficulty of locating the global Φ minimum for any one of the n-mers. For these reasons it has thus far only been practical [16,17] to catalog all stable folding structures for $n<8$. In order to penetrate substantially larger degrees of polymerization n one can focus attention on special families of primary sequence patterns. One such family involves the center-doped molecules ($m>0$)

$$(A)_m-B-(A)_m , \quad (3.1)$$

i.e., a single B residue flanked symmetrically by pure-A tails. We have not examined the inequivalent reverse case with a single A flanked by B tails.

The $m=1$ case A-B-A possesses a single backbone bending degree of freedom and direct minimization [16] shows that the lowest energy is attained when the bend angle $\theta_2=\pm 111.4°$. The potential energy Φ for this optimal shape is listed in Table I.

The search over the $2m-1$ angles for global minima of subsequent family members involved a mixed strategy. In part this strategy utilized a high-temperature Monte Carlo procedure [17] to generate a large, unbiased set of initial configurations for subsequent minimization by both quasi-Newton and conjugate-gradient routines [17,23]. These initial configurations were supplemented by a relatively small, but select, set of configurations created by hand to represent intuitive guesses about likely candidate structures. As it turned out, global minima were redundantly and consistently produced from both sources of initial configurations, thus lending confidence to results reported below.

TABLE I. Potential energies at the global minima for successive center-doped molecules $(A)_m-B-(A)_m$.

m	Φ
0	0.0000
1	−0.6582
2	−2.5317
3	−4.8794
4	−7.7251
5	−11.0670
6	−14.3573
7	−17.9292
8	−21.3945
9	−25.2030
10	−28.6905
11	−32.5054
12	−36.4139

Table I collects the potential energies of the global minima found for the center-doped sequence through $m=12$ ($n=25$); entries of course are expressed in terms of the energy unit selected for the model, the A-A pair interaction depth. The values shown in Table I have a smoothly declining trend, approaching approximately linear behavior with increasing m. First differences $\Phi(m+1)-\Phi(m)$, however, are not monotonic, showing that subtle effects are embedded in the results.

Figures 1–3 illustrate the folded forms of the global minima. The central B residue is shown as a black circle for clarity, the A's as white circles. Not surprisingly, the B dopant in all cases occurs at the exterior of the compact shape, consigned in fact to a relatively remote corner. This is expected since the AB pair interaction is positive for all pair separations, while that for AA pairs is negative beyond unit separation.

We remark in passing that the excitation energies for the center-doped sequence, from global minimum to the next lowest minimum, are on the order of several tenths of an energy unit, but vary considerably and irregularly with m. Even in the cases of small excitation energy (e.g., 0.015 936 for $m=4$), the configurational change involves a substantial shift of a large portion of the molecule.

An important characteristic that is clear from the Figs. 1–3 is a spontaneous symmetry-breaking phenomenon. Nominally the two pure-A tails linked together at the central B residue are indistinguishable. But in the optimally folded forms for all $m>1$ the two tails adopt inequivalent shapes. Since there are two ways of assigning the tails to these different shapes, the global minima must be at least doubly degenerate. Furthermore, each optimal structure is mirror asymmetric, that is, inequivalent

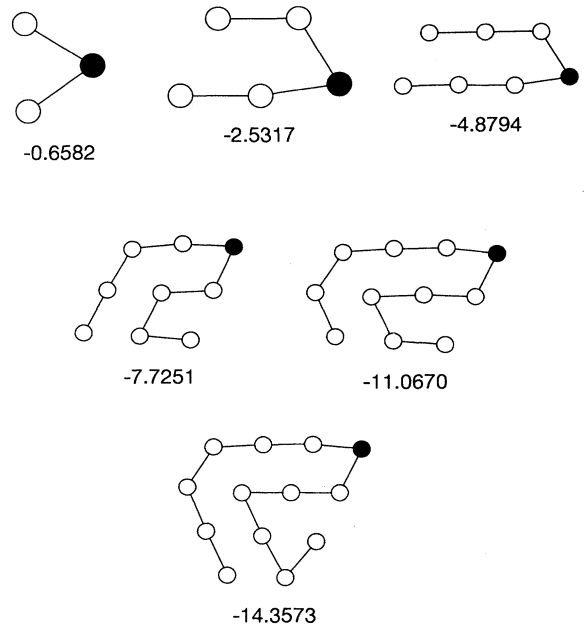

FIG. 1. Structures for the first six members of the center-doped sequence of toy-model polypeptides, each at their global potential energy minima. Open circles represent A units, filled circles represent B units.

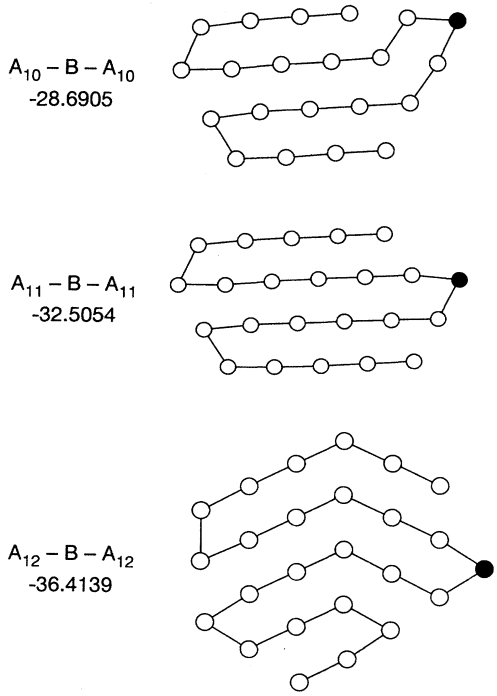

FIG. 3. Global potential energy minimum structures for center-doped molecules with A tails of lengths 10–12.

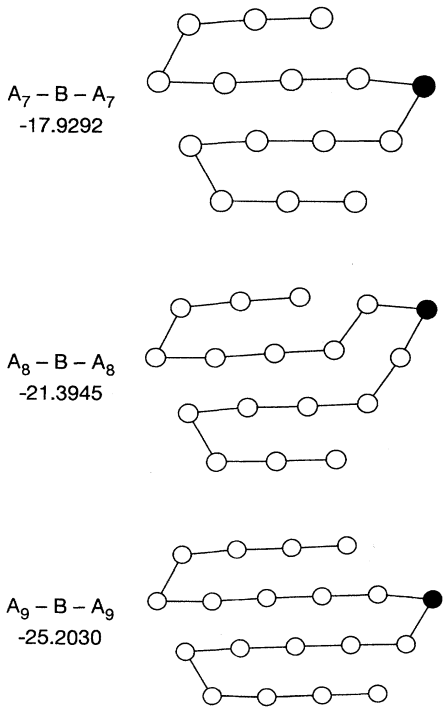

FIG. 2. Global potential energy minimum structures for center-doped molecules with A tails of lengths 7–9.

to its image obtained by reflection across any line in the embedding plane. Therefore each global minimum is fourfold degenerate for $m > 1$, but twofold degenerate in the simple $m = 1$ case.

The mutual arrangement of A residues from both tails exhibits considerable regularity. It seems appropriate to describe the spatial patterns roughly as "crystalline," keeping in mind that the outcome for any m represents a complex compromise of competing effects. On the one hand, the attractive AA pair interactions by themselves would lead to a compact aggregate with maximal internal contact (close packing) and minimal surface. On the other hand, any A aggregate must necessarily be entirely threaded by the backbone of unit-length bonds between successive residues, a "Hamilton chain" in the terminology of graph theory [24]. Furthermore, this Hamilton chain must have its central node, the B dopant, protruding from the exterior of the aggregate. In order to minimize the penalty of backbone bend, the circuit should have as few major direction changes as possible.

A significant disparity exists between the unit bond lengths along the backbone and the distance $r_{ij} = 2^{1/6} = 1.122\,46\ldots$ at which the AA Lennard-Jones attractive pair interaction attains its minimum. This implies that the natural crystal structure for linear and parallel A chains in the plane is distorted from the familiar sixfold symmetric triangular array. Consequently, global minima for the center-doped sequence that display

one tail enfolding a compact arrangement in the other tail (e.g., $m=6$ in Fig. 1) or even of another portion of itself (e.g., $m=12$ in Fig. 3) must involve elastic distortion and elastic energy.

One of the surprises presented by our results was the nonconformity of the $m=12$ case to the few that preceded it. As tail length m increased from 7 to 11 a pattern regularity appeared to emerge, building up an A crystallite row by row, while maintaining the central B dopant at its exterior. Straightforward extrapolation of the $7 \leq m \leq 11$ patterns to $m=12$ would not anticipate the overall exterior shape or the obvious bend across the interior displayed in Fig. 3.

IV. FIBONACCI SEQUENCE

The center-doped sequence discussed in the preceding section represents an extreme of simplicity in its primary structure (residue sequence along the backbone). Its members are nominally symmetric about the backbone midpoint and approach pure-A composition in the limit of large degree of polymerization. We now examine a starkly contrasting sequence.

The arithmetic sequence of Fibonacci numbers $f_0, f_1, f_2, \ldots$ is generated by the specification [25]

$$f_0=0, \quad f_1=1, \quad f_{i+1}=f_{i-1}+f_i \quad (i>0) . \tag{4.1}$$

The magnitudes of the f_i increase rapidly with index i, approaching the asymptotic limiting form

$$f_i \sim 0.447\,213\ldots \gamma^i , \tag{4.2}$$

where γ is the reciprocal of the famous "golden mean" ratio

$$\gamma=(5^{1/2}+1)/2 . \tag{4.3}$$

A simple mapping generates strings of A's and B's from the Fibonacci numbers and we interpret those strings as the members of the Fibonacci sequence of toy model polypeptide molecules. The analog of Eqs. (4.1) is

$$S_0=A, \quad S_1=B, \quad S_{i+1}=S_{i-1}*S_i , \tag{4.4}$$

where now $*$ means concatenation of the literal strings. Following this rule, the leading members of our Fibonacci sequence are found to be

$$A, B, AB, BAB, ABBAB, BABABBAB, \ldots \tag{4.5}$$

Several basic properties can easily be demonstrated for the Fibonacci molecule sequence: (i) all A residues are isolated along the backbone, i.e., flanked on both sides by B's; (ii) B's appear only isolated or in pairs, never as longer uninterrupted B strings; (iii) the limiting fraction of B's for very large strings is $\gamma^{-1}=0.618\,033\ldots$; (iv) the number of residues in S_i is f_{i+1}; (v) the molecules have an hierarchical string structure

$$\begin{aligned} S_i &\equiv S_{i-2}*S_{i-3}*S_{i-2} \\ &\equiv (S_{i-4}*S_{i-5}*S_{i-4})*(S_{i-5}*S_{i-6}*S_{i-5}) \\ &*(S_{i-4}*S_{i-5}*S_{i-4}) \\ &\equiv \cdots . \end{aligned} \tag{4.6}$$

TABLE II. Potential energies at the global minima for leading members of the Fibonacci sequence of molecules.

i	$n(i)$[a]	Φ
0	1	0.0000
1	1	0.0000
2	2	0.0000
3	3	−0.0303
4	5	−0.0057
5	8	−1.4682
6	13	−3.2235
7	21	−5.2881
8	34	−8.9749
9	55	−14.4089

[a]The number of residues $n(i)=f_{i+1}$.

Each of these properties provides a contrast with the center-doped sequence of Sec. III.

The search for global minima was carried out as described above. The resulting potential energies are listed in Table II. While these energies seem to be declining montonically with n, it is not clear what the asymptotic behavior for large n should be.

The geometric structures of the Fibonacci sequence global minima appear in Figs. 4–7. They exhibit far less regularity than the center-doped sequence results, owing to the more nearly equal proportions of A's and B's distributed uniformly along the backbone, with consequent overall interaction frustration.

The hierarchical character of the primary structure, indicated in Eq. (4.6), strongly suggests comparing the geometries of nominally equivalent, embedded, sub-

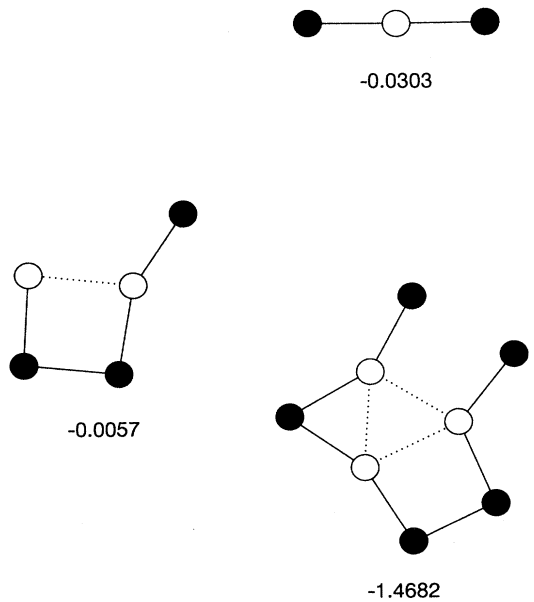

FIG. 4. Structures for the first three members of the Fibonacci sequence. Each molecule is shown at its absolute potential minimum. A and B units are shown, respectively, as open and filled circles. Close AA contacts are indicated by dotted lines.

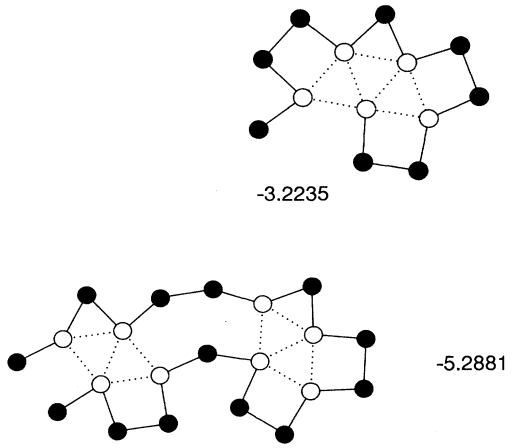

FIG. 5. Global minimum structures for the fourth and fifth members of the Fibonacci sequence, containing 13 and 21 units, respectively.

strings. Following the first of the Eq. (4.6) identities, we can, for example, partition the $n=13$ case as follows:

$$ABBAB*BAB*ABBAB\ . \tag{4.7}$$

Then, by examining the optimized $n=13$ structure in Fig. 5, one sees that the initial and final parameters have roughly similar (but certainly not identical) "dipper" shapes reminiscent of the form adopted by the isolated $n=5$ molecule, Fig. 4.

The next member of the sequence, $n=21$, presents a rather different behavior. The corresponding string division is

$$BABABBAB*ABBAB*BABABBAB\ . \tag{4.8}$$

While the central pentamer again roughly adopts the dipper shape, the two terminal octamers differ substantially from one another: one is relatively extended, the other more compactly folded. This kind of discrepancy between the terminal substring shapes is also evident in the $n=34$ and 55 optimal folding geometries.

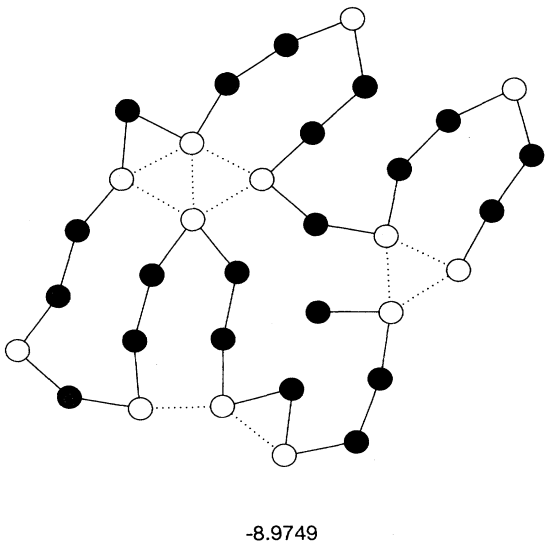

FIG. 6. Global minimum structure for the sixth member of the Fibonacci sequence, containing 34 units.

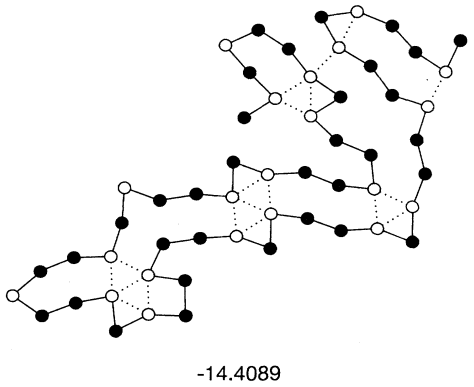

FIG. 7. Global minimum structure for the seventh member of the Fibonacci sequence, containing 55 units.

Close contacts between nonbonded but strongly attracting AA pairs have been highlighted with dotted lines in Figs. 4–7. These of course provide the principal stabilizing elements in the global minimum structures. The intrinsic dispersal of the A's along the Fibonacci sequence backbones means that not all A's can cluster compactly together, even though this by itself would lower the potential energy substantially. Instead, the A's are forced by the backbone connectivity to aggregate into several smaller separated groupings. Predicting how many groupings should appear in the global minimum and which A's they should incorporate does not seem to be possible using elementary concepts.

V. CONCLUSION

The intrinsic difficulty of solving the protein folding problem has been illustrated using a simple (but nontrivial) toy model. Global potential energy minima and their folding structures have been determined for the toy polypeptides forming leading members of two contrasting sequences: center doped and Fibonacci. The results confirm earlier suggestions [16] based on neural network analysis of lower-order (smaller molecule) folding patterns in the model, namely, that primary sequence local information is generally insufficient to predict overall folding geometries. As a result of competing interactions in the toy model (just as in real proteins), collective effects emerge that effectively involve all monomers. The spontaneous symmetry-breaking phenomenon observed in the center-doped sequence is one of these; so too is the unanticipated bend that suddenly appeared across the center of the 25-residue member of this sequence. And in spite of the inherent hierarchical nature of the primary structures in the Fibonacci sequence, competing and frustrated interactions over the entire molecules prevent the natural Fibonacci subsets from consistently adopting the same forms from case to case.

VI. DISCUSSION

The presence and influence of large-scale collective phenomena in toy model protein folding is consistent with the exponential-in-n difficulty that should be expected for minimization problems of this general class. This suggests that an efficient, accurate, and universally applicable algorithm will never be attained to solve all conceivable protein folding problems. Pursuing such an ambitious goal is not necessary however. The preponderance of biological evidence seems to suggest a much more limited objective, namely, the ability to predict the folding outcome for a very small subset of all possible polypeptides, specifically those that reproducibly renature from an unfolded state without substantial kinetic hindrance due to deep metastable traps [8]. Presumably it is only this special subset that can be useful to living organisms. Even within a distinguished "foldable" subset, whether for a toy model or for real polypeptides, a worthy short-term goal would be to improve upon the statistical success rates of the various secondary and tertiary folding structure predictors. This may best proceed by recognizing important collective variables for the protein molecules as a whole, to supplement the local amino acid "window" that has traditionally been employed, particularly in neural network applications [18–22].

In our initial study of the present toy model [16], a complete data base of optimally folded structures was created for several small values of n and then the simplest neural network architecture was identified, which could act as an error-free "predictor" (more precisely, a read-only memory device) for that entire data base. Significantly, these optimal networks contained hidden layers of neurons whose operation was to create collective variables from the primary sequence specification and to supply those collective variables downstream to the output neurons. It strikes us as important generally to try to identify such relevant collective variables, whether for a toy model or for real three-dimensional proteins, without the necessity of solving the very difficult problem of optimizing over all possible neural network architectures.

One possibility that automatically comes to mind are Fourier transform variables. In the present toy model context these would simply be generated from the binary species variable string $\xi_1 \cdots \xi_n$, e.g.,

$$\eta(k) = \sum_{j=1}^{n} \xi_j \exp(ikj) , \tag{6.1}$$

for appropriate k's, with real and imaginary parts supplied as neural network inputs. Alternatively, one might examine Legendre or Chebyshev polynomial [26] transforms as candidate collective variables. Real proteins contain 20 distinct residues rather than just 2, but for them the ξ_i sequence might appropriately be replaced by a numerical hydrophobicity-hydrophilicity index. In any event, the $\eta(k)$ have the requisite nonlocal character for collective folding variables and may even be appropriate for the description of the elastic strain that results above, suggesting that it may be important in real proteins.

ACKNOWLEDGMENTS

T.H.-G. gratefully acknowledges support of the Air Force Office of Sponsored Research, Grant No. F49620-94-1-0081, and the Office of Health and Environmental Research, Department of Energy, under contract No. DE-AC03-76-SF00098.

[1] H. Abe and N. Go, Biopolymers **20**, 1013 (1981).
[2] A. Kolinski, J. Skolnick, and R. Yaris, J. Chem. Phys. **85**, 3585 (1986).
[3] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[4] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[5] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[6] D. A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **89**, 2536 (1992).
[7] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).
[8] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[9] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[10] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[11] M. Sasai and P. G. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[12] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 9029 (1992).
[13] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990).
[14] J. D. Honeycutt and D. Thirumalai, Biopolymers, **32**, 695 (1992).
[15] M. Fukugita, D. Lancaster, and M. G. Mitchard, Proc. Natl. Acad. Sci. U.S.A. **90**, 6365 (1993).
[16] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).
[17] T. Head-Gordon and F. H. Stillinger, Phys. Rev. E **48**, 1502 (1993).
[18] H. Bohr, J. Bohr, S. Brunak, and R. M. J. Cotterill, FEBS Lett. **261**, 43 (1990).
[19] J. D. Hirst and M. J. E. Sternberg, Protein Eng. **4**, 615 (1991).
[20] N. Qian and T. J. Sejnowski, J. Mol. Biol. **202**, 865 (1988).
[21] L. H. Holley and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **86**, 152 (1989).
[22] D. G. Kneller, F. E. Cohen, and R. Langridge, J. Mol. Biol. **214**, 171 (1990).
[23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, New York, 1986), pp. 301–306.
[24] R. G. Busacker and T. L. Saaty, *Finite Graphs and Networks* (McGraw-Hill, New York, 1965), p. 42.
[25] *Applications of Fibonacci Numbers,* edited by G. E. Bergum, A. N. Philippou, and A. F. Horadam (Kluwer Academic, Dordrecht, 1990), Vol. 3.
[26] *Handbook by Mathematical Functions,* Nat. Bur. Stand. Appl. Math. Ser. No. 55, edited by M. Abramowitz and I. A. Stegun (U.S. GPO, Washington, DC, 1968), Chap. 22.