Optimal neural networks for protein-structure prediction

Teresa Head-Gordon
Lawrence Berkeley Laboratories, Berkeley, California 94720

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

(Received 11 January 1993)

The successful application of neural-network algorithms for prediction of protein structure is stymied by three problem areas: the sparsity of the database of known protein structures, poorly devised network architectures which make the input-output mapping opaque, and a global optimization problem in the multiple-minima space of the network variables. We present a simplified polypeptide model residing in two dimensions with only two amino-acid types, A and B, which allows the determination of the global energy structure for all possible sequences of pentamer, hexamer, and heptamer lengths. This model simplicity allows us to compile a complete structural database and to devise neural networks that reproduce the tertiary structure of all sequences with absolute accuracy and with the smallest number of network variables. These optimal networks reveal that the three problem areas are convoluted, but that thoughtful network designs can actually deconvolute these detrimental traits to provide network algorithms that genuinely impact on the ability of the network to generalize or learn the desired mappings. Furthermore, the two-dimensional polypeptide model shows sufficient chemical complexity so that transfer of neural-network technology to more realistic three-dimensional proteins is evident.

PACS number(s): 87.10.+e

I. INTRODUCTION

The application of neural networks as a computational tool is rapidly expanding into the physical sciences [1], engineering [2], mathematics [3], and even managerial [4] studies. However, their successful use is often hampered by the lack of rigorous proofs that specific network architectures (neuronal connections and biases) are optimally designed for a given pattern-recognition or memory-recall task. The utilization of neural networks for direct predictions of protein structure [5–20] provides an example of this difficulty. The objective in this case is to convert the information about the primary structure (the amino-acid sequence) into predictions about the secondary structure [5–15] (local chain-folding preferences) and tertiary structure [15–20] (overall protein-folding pattern). Until recently, the best network designs for secondary-structure prediction did somewhat better than other non-network statistical methods [20–24], but did not seem able to improve beyond an average of 65% overall predictive capacity. Recently reported statistical methods [25,26] are now in fact competing effectively with ~65% prediction accuracy. Furthermore, certain types of secondary structure are predicted much less well than this average, such as β turns, where the best network predictions were 26% [11]. The prediction of tertiary structure [15–20] has also been tried, but with limited success when compared to sequence-homology methods [15].

Three possible sources of error exist for neural-network prediction of protein structure. First, the experimental database of known protein structures is extremely sparse compared to the entire family of possible proteins with comparable degree of polymerization; this alone suggests that network training using some or all of the database would be insufficient. Second, the network topologies themselves may not permit effective *learning* strategies, so that the network is unable to adaptively predict, or generalize to, a new data set of sequence-structure relationships. Finally, neural networks from a mathematical-optimization standpoint are known to suffer from their own multiple-minimum problem, so that optimal solutions are not easily attainable.

Furthermore, these three fundamental problems, which are in themselves quite formidable, further exacerbate poor neural-network predictive capacity by their convolution. Database sparsity is thought to be responsible for the observation that hidden layers do not improve secondary structure predictions because there is simply not enough higher-order information (representative interactions between two or more amino acids) to exploit the full power of such neural-network topologies [13,14]. In this case, neural networks would be confounded by the distinction between sequences for which a specific substitution (mutation) leaves the folding pattern unchanged, from those that experience profound folding changes resulting from the same specific substitution. Information that is present in the database may be lost for neural-network topologies with too many free parameters and arbitrary training criteria. The network in this case has simply "fit" the data in a nonlinear least-squares sense so that the network performs exceptionally well on the training database, but is overtrained so that generalization to the testing set is impossible [19]. The multiple minima problem, whereby "converged" neural network weight and bias parameters define a *local* minimum, results in the trapping of the training process into network solution minima which are not optimal [10].

The present work tries to address the three stated inadequacies, and their deconvolution, by designing networks which perform perfectly for a database complete up to a given sequence size that describes small, two-dimensional "polypeptides," and to infer what neural-network topologies may be required for longer polypeptides in three dimensions with full sequence diversity. Needless to say, the prediction of protein secondary and tertiary structures is a highly complex problem. With 20 commonly occurring amino acids and polypeptide sizes ranging from 10^2 to 10^3 residues, the diversity of plausible protein structures is vast. Because the interactions that give rise to the native tertiary structure are not well understood, and which in fact define the problem of protein folding, we have chosen to strip away a great deal of this complexity in order to tackle best the problem of optimal neural networks. The two-dimensional model polypeptides that we have studied are composed of only two amino-acid types, and the size regimes we have considered thus far extend from trimers up to heptamers. This simplification permits several advantages. First, we are able to enumerate and specify structures for the full sequence space for the polypeptide sizes considered here. This is certainly not plausible theoretically, let alone experimentally, for polypeptides of greater lengths composed of 20 or more amino acid types. Second, we have a much better chance of identifying the global energy minimum for each of these small two-dimensional polypeptides; in three dimensions the global energy minimum is thought usually to be synonymous with the native structure. We present what we hope are convincing arguments that we have found the "native" structures for all amino-acid sequences and for all polypeptide lengths considered. Having compiled a complete database we are then free to design networks which reproduce perfectly a structural feature of interest, internal coordinates, or residue-residue contacts. We define optimal networks as those with a minimal number of network variables (weights and biases) which predict the structural database exactly. We emphasize at this point that, although the model polypeptides we use are quite simple, they show a wide variety of native structures which may be described loosely in terms of the three-dimensional protein structural categories of linear, sheet, helix, and globular once polypeptide lengths of at least pentamer size are reached. Thus in this paper we focus strictly on the length regime of pentamers through heptamers.

In Sec. II we present the two-dimensional model potential-energy function, the procedure used for finding global energy-minimum structures, and the protocols used to eliminate symmetries in the sequences and global energy minimum structures. Section III describes the neural-network architectures which we believe optimally reproduce the residue-residue contacts of all amino-acid sequences for a given polypeptide length ranging from pentamer to heptamer. We provide a discussion in Sec. IV of what is required for extending perfect network designs to more realistic protein systems and ideas currently being pursued.

II. MODEL DESCRIPTIONS

A. Potential-energy function

Our model polypeptides consist of linear strings of structureless "amino-acid" monomers that can be either of two types, A and B. The bonds connecting neighboring monomers have fixed unit length, but successive bonds can change direction by a bend degree of freedom at each nonterminal monomer. This family of model molecules resides in two-dimensional Cartesian space.

The potential-energy function for the general n-mer polypeptide is

$$\Phi = K \sum_{j=2}^{n-1} [1-\cos(\theta_j)] + 4 \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} [r_{ij}^{-12} + f(\zeta_i,\zeta_j) r_{ij}^{-6}] \ . \tag{2.1}$$

Here the amino-acid monomers have been numbered sequentially along the bond backbone. Angle θ_j measures the bend away from linearity for the two bonds impinging on monomer j. The distance between monomers i and j has been denoted by r_{ij}. Monomer species are indicated by parameters $\zeta_1, \ldots, \zeta_n$, with value $+1$ for A and -1 for B. The function $f(\zeta_i,\zeta_j)$ may be written as follows:

$$f(\zeta_i,\zeta_j) = \tfrac{1}{2} - \tfrac{1}{8}(\zeta_i+\zeta_j) - \tfrac{5}{16}(\zeta_i+\zeta_j)^2 \ ; \tag{2.2}$$

it equals -1 for an AA pair, $-\frac{1}{2}$ for a BB pair, and $+\frac{1}{2}$ for an AB pair. In all of the calculations reported below the bend force constant K has been assigned the value $\frac{1}{4}$.

For the purpose of numerical computation it is useful to define a laboratory fixed coordinate system to measure orientation angles for the bonds. Let these be denoted by $\alpha_1, \ldots, \alpha_{n-1}$, as shown in Fig. 1. By convention counterclockwise rotations will be taken as positive. If amino acid 1 is assumed to remain at the origin, then the Cartesian coordinates of the other $n-1$ particles are given by

$$x_i = \sum_{j=1}^{i-1} \cos(\alpha_j) \ , \quad y_i = \sum_{j=1}^{i-1} \sin(\alpha_j) \ . \tag{2.3}$$

Also

$$r_{ij}^2 = \left[\sum_{k=i}^{j-1} \cos(\alpha_k) \right]^2 + \left[\sum_{k=i}^{j-1} \sin(\alpha_k) \right]^2 \tag{2.4}$$

and

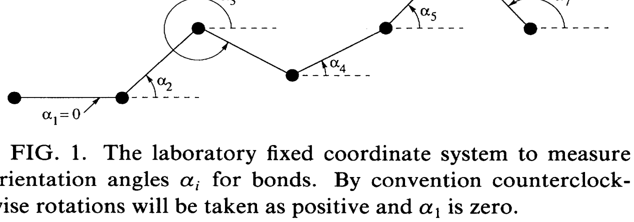

FIG. 1. The laboratory fixed coordinate system to measure orientation angles α_i for bonds. By convention counterclockwise rotations will be taken as positive and α_1 is zero.

$$\theta_i = \alpha_{i-1} - \alpha_i \quad (2 \le i \le n-1) . \tag{2.5}$$

We adhere to the convention displayed in Fig. 1 where $\alpha_1 = 0$. Derivatives with respect to orientation angles α can be easily obtained for use in minimization algorithms.

B. Amino-acid sequences and protocols

While native polypeptides have natural directionality by the fact that they have an N terminus and a C terminus, our model peptides have no such directionality; sequence AABB, for example, is the same for all purposes as BBAA. Although these sequence degeneracies can be lifted by placing groups C (N terminus) and D (C terminus) on the polypeptide ends, such a scheme would add unwanted complexity to our peptide model at this point. We instead have chosen the following convention for eliminating symmetries in sequence space: directionality is defined by taking the sequence with the larger string of A's as scanned from left to right (i.e., alphabetical order). Thus, in the example given above, only AABB is retained while BBAA is eliminated from consideration. Formally for any polypeptide of length n, we can then enumerate all possible remaining sequences: for the case of the ($2m$)-mer we have $2^{m-1}(2^m+1)$ possible sequences, while there are $2^m(2^m+1)$ possible sequences for the ($2m+1$)-mer.

In addition, a second convention is required in order to decide among structural alternatives when geometric multiplicities are found for the global energy minimum of certain sequences. For example, Fig. 2 shows two energetically equivalent global energy structures for the amino acid sequence AABBAA; these structures are mirror images and only differ in their specific contacts, σ_{ij}. In such cases, we choose one among the n-fold energy degenerates by invoking the convention that the conformer which minimizes the sum of contact labels i and j is the desired structure:

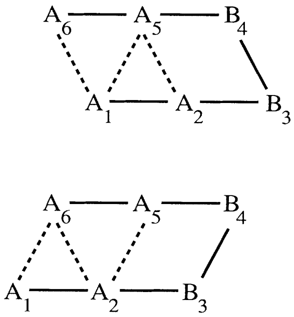

FIG. 2. Two energetically equivalent global energy structures for the amino-acid sequence AABBAA; these structures are mirror images and only differ in their specific contacts, σ_{ij}. In such cases, we choose one among the n-fold energy degenerates by invoking the convention that the conformer which minimizes the sum of $+1$ contact labels i and j is the desired structure.

$$\min \sum_{i=1}^{n} \sum_{j=i+1}^{n} \tfrac{1}{2}[(i+j)(\sigma_{ij}+1)] \tag{2.6}$$

where n is the number of residues and $\sigma_{ij}=1$ or -1 for

TABLE I. Sequence and structural database for 2D pentamer. All sequences studied are given in the first column. The columns denoted by α_i together specify the orientation angles in a laboratory fixed frame for the global energy structure of each sequence, where $\alpha_1=0$ by convention (see Fig. 1). The columns denoted by σ_{jk} indicate the nonbonded contacts which are present ($+1$) and those which are not (-1) for the global energy structure of each sequence.

Sequence	α_2 (deg)	α_3 (deg)	α_4 (deg)	σ_{13}	σ_{14}	σ_{15}	σ_{24}	σ_{25}	σ_{35}
AAAAA	8	120	180	−1	−1	1	1	1	−1
AAAAB	112	172	160	1	1	−1	−1	−1	−1
AAABA	54	113	225	−1	−1	1	−1	1	1
AAABB	111	102	101	1	−1	−1	−1	−1	−1
AABAA	8	120	180	−1	−1	1	1	1	−1
AABAB	60	172	74	−1	1	−1	1	−1	1
AABBA	30	117	202	−1	−1	1	−1	1	−1
AABBB	0	0	0	−1	−1	−1	−1	−1	−1
ABAAB	112	172	160	1	1	−1	−1	−1	−1
ABABA	111	120	231	1	−1	1	−1	−1	1
ABABB	111	98	98	1	−1	−1	−1	−1	−1
ABBAB	86	171	146	−1	1	−1	−1	−1	−1
ABBBA	44	144	189	−1	−1	1	1	−1	−1
ABBBB	350	52	153	−1	−1	−1	−1	1	1
BAAAB	353	104	97	−1	−1	−1	1	−1	−1
BAABB	0	0	0	−1	−1	−1	−1	−1	−1
BABAB	350	101	92	−1	−1	−1	1	−1	−1
BABBB	103	163	211	1	1	1	−1	−1	−1
BBABB	18	121	179	−1	−1	1	1	1	−1
BBBBB	17	119	180	−1	−1	1	1	1	−1

contact or noncontact.

C. Global energy search procedure

The potential function described above is simple enough so that Monte Carlo searches can be reasonably exhaustive and also chemical "intuition" for the structure optimality can be invoked. We have used both procedures to determine the global energy minimum for each sequence of each polypeptide length $n \leq 7$. The Monte Carlo and minimization procedure involved heating a given n-mer sequence to reduced temperature 2×10^4 and sampling configurations every 1×10^4 steps during a 5×10^4 step run. The resulting 50 configurations were minimized by a Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimization algorithm. To provide a check on the Monte Carlo search results, we also minimized alternative structures derived from chemical intuition, where the possibility of a larger number of favorable contacts was found or where angle strain was not as great as that found by the Monte Carlo procedure. Rarely did we find structures in this way which were lower in energy than that found with the Monte Carlo–minimization scheme. Tables I–III lists the sequences investigated by the global energy search described above (see convention definitions defined above) and the global minimum bond angles α_i and contacts σ_{ij} for polypeptide sizes $n = 5$, 6, and 7.

D. Structural diversity displayed by peptide model

It is obvious from Eq. (2.1) that Φ includes two types of interactions, backbone bend energy and Lennard-Jones interactions between nonbonded monomers. These types come into conflict in establishing the global energy minimum for any sequence; only by bending the backbone is it possible to attain substantial energy lowering through attractive nonbonded pair potentials. Changing the sequence of A's and B's shifts the balance between competing Φ contributions and can produce a diversity of minimum-Φ structures.

Figures 3–6 provide global energy structures for different sequences which are representative of the

TABLE II. Sequence and structural database for 2D hexamer. See Table I caption.

Sequence	α_2 (deg)	α_3 (deg)	α_4 (deg)	α_5 (deg)	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{24}	σ_{25}	σ_{26}	σ_{35}	σ_{36}	σ_{46}
AAAAAA	0	59	171	179	−1	−1	−1	1	−1	1	1	1	−1	−1
AAAAAB	60	172	180	168	−1	1	1	−1	1	−1	−1	−1	−1	−1
AAAABA	112	172	173	285	1	1	−1	1	−1	−1	−1	−1	−1	1
AAABAA	1	61	173	181	−1	−1	−1	1	−1	1	1	1	−1	−1
AAAABB	111	172	161	161	1	1	−1	−1	−1	−1	−1	−1	−1	−1
AAABAB	53	113	225	124	−1	−1	1	−1	−1	1	−1	1	−1	1
AAABBA	109	92	176	265	1	−1	−1	1	−1	−1	−1	−1	1	−1
AABAAB	60	172	180	164	−1	1	1	−1	1	−1	−1	−1	−1	−1
ABAAAB	112	171	225	210	1	1	1	−1	−1	−1	−1	−1	−1	−1
BAAAAB	340	91	151	170	−1	−1	−1	1	1	1	−1	−1	−1	−1
AABABA	1	113	121	232	−1	−1	−1	1	1	−1	−1	1	−1	1
ABAABA	112	114	173	285	1	1	−1	1	−1	−1	−1	−1	−1	1
AABBAA	29	117	201	181	−1	−1	1	1	−1	1	−1	−1	−1	−1
AAABBB	111	98	95	96	1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AABABB	60	172	69	12	−1	1	−1	−1	1	−1	−1	1	1	−1
ABAABB	112	171	161	161	1	1	−1	−1	−1	−1	−1	−1	−1	−1
BAAABB	28	139	170	197	−1	−1	1	1	1	−1	−1	−1	−1	−1
AABBAB	29	117	201	100	−1	−1	1	−1	−1	1	−1	−1	−1	1
ABABAB	112	120	231	180	1	−1	1	−1	−1	−1	−1	1	−1	1
BAABAB	19	79	190	169	−1	−1	−1	1	−1	1	−1	1	−1	−1
AABBBA	17	67	164	208	−1	−1	−1	1	−1	−1	1	1	−1	−1
ABABBA	111	92	177	264	1	−1	−1	1	−1	−1	−1	−1	1	−1
BAABBA	348	18	105	190	−1	−1	−1	−1	−1	−1	1	−1	1	−1
AABBBB	2	353	54	154	−1	−1	−1	−1	−1	−1	−1	−1	1	1
ABABBB	111	10	309	262	1	−1	−1	−1	1	1	1	−1	−1	−1
BAABBB	3	3	4	3	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABBABB	17	13	270	208	−1	−1	−1	−1	−1	−1	−1	1	1	−1
ABBBAB	45	144	189	157	−1	−1	1	−1	1	−1	−1	−1	−1	−1
ABBBBA	1	99	164	190	−1	−1	−1	1	1	1	−1	−1	−1	−1
BABABB	333	11	318	290	−1	−1	1	1	1	−1	−1	−1	−1	−1
BABBAB	0	86	171	171	−1	−1	−1	1	−1	1	−1	−1	−1	−1
BBAABB	0	0	0	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABBBBB	349	4	105	168	−1	−1	−1	−1	−1	−1	1	1	1	−1
BABBBB	102	161	212	254	1	1	1	1	−1	−1	−1	−1	−1	−1
BBABBB	16	120	178	182	−1	−1	1	1	1	1	−1	−1	−1	−1
BBBBBB	16	117	180	181	−1	−1	1	1	1	1	−1	−1	−1	−1

TABLE III. Sequence and structural-database for 2D heptamer. See Table I caption.

Sequence	α_2 (deg)	α_3 (deg)	α_4 (deg)	α_5 (deg)	α_6 (deg)	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{17}	σ_{24}	σ_{25}	σ_{26}	σ_{27}	σ_{35}	σ_{36}	σ_{37}	σ_{46}	σ_{47}	σ_{57}
AAAAAAA	107	353	293	239	184	1	−1	−1	−1	1	1	1	1	1	−1	−1	−1	−1	−1	−1
AAAAAAB	352	241	180	180	192	−1	−1	1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1	−1
AAAAABA	61	172	180	181	293	−1	1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	1
AAAABAA	307	306	247	134	127	−1	−1	−1	−1	1	−1	−1	−1	1	−1	1	1	1	−1	−1
AAAAABB	61	172	180	169	168	−1	1	1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AAAABAB	359	248	239	128	230	−1	−1	−1	1	−1	1	−1	1	−1	−1	−1	−1	1	−1	1
AAAABBA	53	105	134	222	306	−1	−1	−1	−1	1	−1	−1	−1	1	−1	−1	1	−1	1	−1
AAAABBB	249	188	201	197	194	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AAABAAA	0	352	240	181	180	−1	−1	−1	−1	1	−1	−1	1	1	1	1	−1	−1	−1	−1
AAABAAB	0	300	188	180	224	−1	−1	−1	1	−1	−1	1	1	−1	1	−1	−1	−1	−1	−1
AAABABA	308	307	195	187	75	−1	−1	−1	−1	1	−1	−1	−1	1	1	−1	1	−1	−1	1
AAABBAA	0	331	244	159	179	−1	−1	−1	−1	1	−1	−1	1	1	−1	1	−1	−1	−1	−1
AAABABB	53	113	225	122	64	−1	−1	1	−1	−1	−1	1	−1	−1	1	−1	−1	1	1	−1
AAABBAB	249	257	255	252	262	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AAABBBA	52	74	119	219	263	−1	−1	−1	−1	1	−1	−1	−1	1	−1	−1	1	1	−1	−1
AAABBBB	111	103	119	59	316	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	1
AABAAAB	8	120	180	181	180	−1	−1	1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1	−1
AABAABA	60	172	180	180	292	−1	1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	1
AABABAA	50	299	291	179	179	−1	−1	−1	−1	1	1	−1	1	1	−1	−1	−1	1	−1	−1
AABAABB	60	172	180	167	166	−1	1	1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AABABAB	0	248	240	129	230	−1	−1	−1	1	−1	1	−1	1	−1	−1	−1	−1	1	−1	1
AABABBA	0	112	92	176	263	−1	−1	−1	−1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1
AABABBB	60	172	70	9	322	−1	1	−1	−1	−1	1	−1	−1	−1	1	1	1	−1	−1	−1
ABAAAAB	112	172	226	277	264	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABAAABA	248	188	135	134	22	1	1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
ABAAABB	111	172	225	214	214	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABAABAB	112	172	173	285	187	1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	1
ABAABBA	112	113	142	229	314	1	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	−1	1	−1
ABAABBB	248	189	201	202	206	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
BAAAAAB	17	26	137	197	169	−1	−1	−1	−1	1	−1	−1	1	−1	1	1	−1	−1	−1	−1
BAAAABB	18	79	190	170	196	−1	−1	−1	1	1	−1	1	−1	−1	1	−1	−1	−1	−1	−1
BAAABAB	6	313	253	141	239	−1	−1	−1	−1	−1	−1	−1	1	−1	−1	1	−1	1	−1	1
BAAABBB	331	221	191	162	117	−1	−1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1
AABBAAB	331	243	159	179	192	−1	−1	1	1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1
AABBABA	333	58	145	123	234	−1	−1	−1	−1	1	−1	1	−1	1	−1	−1	−1	−1	−1	1
AABBBAA	26	345	244	198	179	−1	−1	−1	−1	1	−1	−1	1	1	1	−1	−1	−1	−1	−1
AABBABB	30	117	202	98	40	−1	−1	1	−1	−1	−1	1	−1	−1	−1	−1	−1	1	1	−1
AABBBAB	19	66	166	208	108	−1	−1	−1	1	−1	−1	−1	1	−1	1	−1	−1	−1	−1	1
AABBBBA	19	21	120	183	209	−1	−1	−1	−1	1	−1	−1	−1	1	1	1	−1	−1	−1	−1
AABBBBB	6	349	3	106	167	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	−1
ABABAAB	248	240	129	127	131	1	−1	1	1	−1	−1	−1	−1	−1	1	−1	−1	−1	−1	−1
ABABABA	112	61	173	181	292	1	−1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1	−1	1
ABABABB	111	120	231	128	70	1	−1	1	−1	−1	−1	−1	−1	−1	1	−1	−1	1	1	−1
ABABBAB	248	270	184	99	175	1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	−1	−1	−1	−1
ABABBBA	248	276	234	134	88	1	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	1	−1	−1
ABABBBB	249	351	11	109	116	1	−1	−1	−1	−1	1	−1	1	1	−1	−1	−1	1	−1	−1
BAABAAB	332	32	144	151	169	−1	−1	−1	−1	1	−1	−1	1	−1	1	1	−1	−1	−1	−1
BAABABB	340	281	169	188	163	−1	−1	−1	1	1	−1	1	−1	−1	1	−1	−1	−1	−1	−1
BAABBAB	17	348	260	176	277	−1	−1	−1	−1	−1	−1	−1	1	−1	−1	1	−1	−1	−1	1
BAABBBB	3	6	348	51	151	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	1
ABBAAAB	86	172	203	254	246	−1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABBAABB	86	173	203	190	190	−1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABBABAB	92	172	159	267	189	−1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	−1
ABBABBA	353	48	149	168	190	−1	−1	−1	−1	1	−1	1	1	−1	1	−1	−1	−1	−1	−1
ABBABBB	324	336	78	136	146	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1
BABAABB	21	270	210	189	160	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1
BABABAB	330	81	89	198	175	−1	−1	−1	−1	1	1	−1	1	−1	−1	−1	−1	1	−1	−1
BABABBB	30	141	174	198	246	−1	−1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1

TABLE III. (*Continued.*)

Sequence	α_2 (deg)	α_3 (deg)	α_4 (deg)	α_5 (deg)	α_6 (deg)	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{17}	σ_{24}	σ_{25}	σ_{26}	σ_{27}	σ_{35}	σ_{36}	σ_{37}	σ_{46}	σ_{47}	σ_{57}
BBAAABB	348	15	126	158	184	−1	−1	−1	−1	1	−1	−1	1	1	1	−1	−1	−1	−1	−1
BBAABBB	10	292	210	175	180	−1	−1	−1	1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1
ABBBAAB	48	146	190	213	205	−1	−1	1	1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1
ABBBABB	350	352	52	153	172	−1	−1	−1	1−	−1	−1	−1	−1	1	−1	1	1	1	−1	−1
ABBBBAB	2	100	165	189	88	−1	−1	−1	1	−1	1	1	−1	−1	−1	−1	−1	−1	−1	1
ABBBBBA	25	41	143	202	190	−1	−1	−1	−1	1	−1	−1	1	−1	1	1	−1	−1	−1	−1
ABBBBBB	11	12	309	206	193	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	1	1	1	−1	−1
BABBABB	101	103	119	223	281	1	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	1	1	−1
BABBBAB	8	325	224	181	189	−1	−1	−1	−1	1	−1	−1	1	−1	1	−1	−1	−1	−1	−1
BBABBBB	103	162	212	257	302	1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
BBABABB	347	15	126	158	182	−1	−1	−1	−1	1	−1	−1	1	1	1	−1	−1	−1	−1	−1
BBABBBB	346	242	183	177	134	−1	−1	1	1	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1
BBBABBB	1	345	241	184	178	−1	−1	−1	−1	1	−1	−1	1	1	1	1	−1	−1	−1	−1
BBBBBBB	359	62	163	180	180	−1	−1	−1	1	1	−1	1	1	−1	1	−1	−1	−1	−1	−1

structural diversity present in our peptide model. Using the hexamer polypeptide length as an example, the lowest-energy conformer for sequence BBAABB is linear (Fig. 3), sequence AAABAA has as its lowest energy structure a "β-sheet" motif (Fig. 4), sequence AABABB may be classified as "helical" (Fig. 5), while sequence AAABAB provides an example of a lowest-energy structure which is globulelike (Fig. 6). These representative structures reveal more clearly the roles of amino-acid types A and B in mixed sequences using the parlance of protein chemistry. Due to the stronger A-A attraction as compared to B-B, residue type A is more likely to form secondary structure "hydrogen bonds" (such as the β-sheet structure in Fig. 4 and the helix in Fig. 5) or to be more hydrophobic (by "maximizing" a "hydrophobic core" as in Fig. 6). Due to its comparatively weaker interaction, amino-acid type B is found to be analogous to the glycine residue in that it is usually found at a turn (Fig. 4) or on the polypeptide exterior (Fig. 6) in mixed sequences. The reasonable structural diversity present in our simple polypeptide model and the associations which exist with real protein chemistry provide sufficient impetus for defining neural-network architectures which perform perfectly for these small model systems in order to provide insight into models for larger and more realistic polypeptide models.

III. NEUTRAL-NETWORK DESIGNS

Neural-network approaches for performing learning functions such as pattern recognition are motivated by the fact that the central nervous system is known to excel at such tasks [1]. In application to the protein-folding problem, neural-network algorithms are required to predict patterns of local and global chain-folding preferences of the native protein (neuronal output) from the amino-acid sequence (input to the network). The common topology of neural networks often used to predict these conformational preferences is known as feedforward-back propagation networks with or without hidden layers [1] (Fig. 7). In this case, each amino acid of a protein sequence is represented by a small set of input channels which are directly connected, or fed into, hidden neurons which in turn connect to output neuron(s) representing a secondary or tertiary-structure classification. The input channels generally correspond to both the amino acid whose most likely secondary or tertiary structure is being predicted, while the remainder supply a context (or window) of n amino acids preceding and succeeding this amino acid along the backbone. The learning, or training, phase of the neural-network algorithm involves minimizing the function

$$E=\sum_{i=1}^{N}\sum_{j=1}^{M}(O_{oj}^{i}-O_{cj}^{i})^2 \qquad (3.1)$$

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

$$A_{cj}^{i}=\sum_{k=1}^{L} w_{jk} I_k^i + b_j \qquad (3.2)$$

with an output response of

$$O_{cj}^{i}=1/[1+\exp(A_{cj}^{i})] \ . \qquad (3.3)$$

In this study, we have used the discontinuous response function

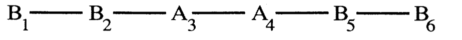

FIG. 3. The lowest-energy conformer for sequence BBAABB is a linear structure.

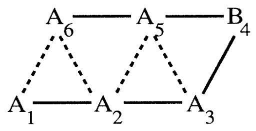

FIG. 4. The lowest-energy conformer for sequence AAABAA is a "β-sheet" motif.

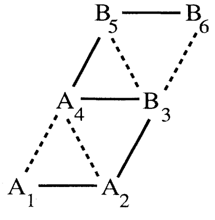

FIG. 5. The lowest-energy conformer of sequence AABABB may be classified as "helical."

$$O^i_{cj}=\text{sgn}(A^i_{cj}) \tag{3.4}$$

where L is the number of input units, I_k is the input, w_{jk} is the weight of the connection between the upstream neuron k, and the downstream neuron j, and b_j is the bias associated with the output neuron j. When a continuous response function such as Eq. (3.3) is employed, a steepest-descent algorithm is often used for minimizing the function in Eq. (3.1) with respect to the free parameters w_{jk} and b_j. The parameters w_{jk} and b_j are updated (or "backpropagated" through the network from output to input) by the following derivative expressions:

$$\Delta w_{jk}=-\gamma\frac{\partial E}{\partial w_{jk}}\ , \tag{3.5}$$

$$\Delta b_j=-\gamma\frac{\partial E}{\partial b_j} \tag{3.6}$$

where γ is a damping, or "learning," factor.

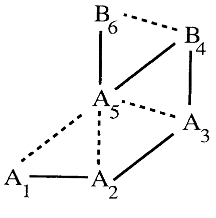

FIG. 6. The lowest-energy conformer for sequence AAABAB provides an example of a globular structure.

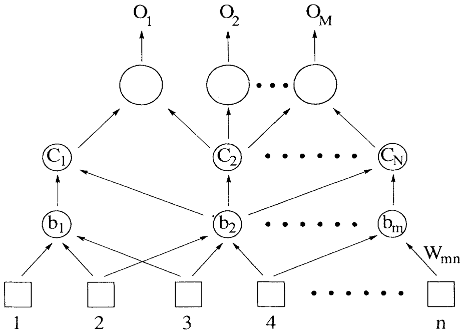

FIG. 7. A generic feedforward neural-network architecture.

In previous applications of feedforward neural networks for predicting secondary [5–15] and tertiary [15–20] protein structure, the sources of error described in the Introduction may have contributed to their diminished predictive accuracy: database incompleteness, nonoptimal network topologies, and mathematical optimization problems. Our approach, involving a much simpler chemical model, will ultimately allow us to deconvolute these debilitating attributes in order to assess their impact on optimal neural networks for predicting three-dimensional proteins with full sequence diversity. As described in Sec. II, we have eliminated the problems of database degradation altogether by finding the global energy structures to the limit of numerical precision for all possible sequences of polypeptides composed of two amino-acid types with lengths ranging from trimer to heptamer. In this section, we will demonstrate the existence of optimal topologies which specify whether two residues are in contact (a distance of 1.34 times backbone bond length, or less) for the full sequence database for each polypeptide length. The perfect prediction of the binary amino-acid contact matrix for each sequence in turn defines the tertiary structure perfectly.

We design network topologies which reproduce with complete accuracy whether two amino acids are ($\sigma_{ij}=+1$) or are not ($\sigma_{ij}=-1$) in contact for all possible sequences for a given individual amino-acid pair. This is accomplished, individually, for all possible nonbonded i,j pairs, where $i+2\leq j\leq n$. This approach is to be contrasted with the simultaneous solution of *all* contact pairs for all sequences. The latter method likely involves a nonintuitive network architecture which can only be found by exhaustive Monte Carlo searches. The parameter spaces searched in this latter case would involve two multiple minima problems: (1) the determination of the optimal number of hidden layers and number of hidden neurons within a layer, and (2) given a certain number of hidden layers and neurons, the search for the optimal solution using a starting architecture in which all possible connections (input→ hidden, hidden → output, and input → output) are present. The benefit of the former tactic is that network topologies can be designed "by hand" with specific Boolean functions [1], which we discuss below. In this case, given an appropriate Boolean function, it is easy to derive the remaining architecture to optimally reproduce the observed contacts. The multiple minimum problem then reduces to a search for the best Boolean function which, together with its remaining architecture, reproduces the observations with the smallest number of network variables with absolute accuracy. Therefore we have avoided some aspects of the mathematical optimization problems usually encountered in conventional applications of neutral networks to proteins structure prediction. Furthermore, our hand-designed architectures make obvious the network topologies which may successfully be applied to neural-network predictions of tertiary structure for proteins with the usual sequence diversity.

We have considered three network architecture types to predict all individual contact values for all possible sequences of the pentamer, hexamer, and heptamer polypeptide lengths. They are distinguished from one and other by a Boolean function which initially differentiates between on $(+1)$ and off (-1) contact values, using the discontinuous response function in Eq. (3.4). These functions are presented in Figs. 8–10 using the notation convention in Ref. [1]. In Figs. 8 and 9, the Boolean function involves the direct connections between inputs *i* and *j* with the output neutron *k*, the latter which describes the value of contacts *i,j*. Figures 8 and 9 differ only in the assigned weights w_{ij} and w_{jk}, which are both $+1$ in Fig. 8 and both -1 in Fig. 9. Thus the network defined by the Boolean function in Fig. 8 (Fig. 9) yields a $A^i_{ck}=+2$ value if the input neurons of *i* and *j* are both A (B) residues, $A^i_{ck}=0$ if the input neurons are A and B or B and A, and $A^i_{ck}=-2$ if both input neurons are B (A) amino acids. Figure 10 corresponds to a Boolean function that discourages predictions of favorable A and B or B and A contacts ($A^i_{ck}=-4$), but encourages like-like interactions (AA *and* BB) with a value of $A^i_{ck}=0$. The use of a discontinuous output response function with a -1 bias value would give a $+1$ output value for Fig. 8 (Fig. 9) for AA (BB) interactions, and -1 value otherwise; for the central Boolean function described in Fig. 10, a discontinuous response function with a bias of $+1$ would predict all AA and BB favorably and disallow AB and BA contacts. However, our interaction potential described in Eqs. (2.1) and (2.2) does not produce such a simple relation between sequence and structure as these individual Boolean functions would alone indicate. In the case of network types described by the Boolean functions in Fig. 8 (Fig. 9), not all A_iA_j (B_iB_j) contacts are favored, nor are all B_iB_j (A_iA_j) contacts forbidden. The remaining network architecture must then turn on *and* turn off those sequences whose contact values are not predicted correctly by that central Boolean function. For networks described by the Boolean function in Fig. 10, not all A_iA_j and B_iB_j pairs are favored, so that the remaining architecture in this case must (only) turn off those sequences where A_iA_j or B_iB_j are not in contact. The benefit of these Boolean function types is that they correctly categorize a large portion of the sequences into on and off contact values using only two or five weights and at most one hidden neuron. Additional architecture is then required to correct the individual Boolean function results on a much smaller subset of the sequence database for a given contact value. For notational convenience we will refer to the network type in Fig. 8 as NN1, that in Fig. 9 as NN2, and that in Fig. 10 as NN3. Table IV provides an explicit architecture for the optimal network solution for the there network types for the full sequence database of the heptamer σ_{13} contact predictions.

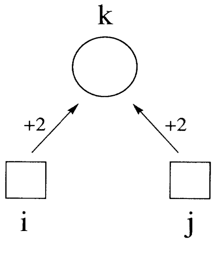

FIG. 8. Central Boolean function for neural network 1.

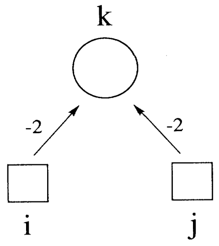

FIG. 9. Central Boolean function for neural network 2.

Tables VIII–XXXV, provided as supplementary material [33] or obtainable from the authors, supply optimal network solutions for the three network types for the full sequence database for all contacts for polypeptide sizes ranging from pentamer to heptamer.

Tables V–VII incorporate a synopsis of the number of network variables required to reproduce the entire contact database for each contact σ_{ij} of the three network architecture types—NN1, NN2, and NN3, for the pentamer, hexamer, and heptamer, respectively. The next to last entry row indicates the number of variables which should be subtracted from the sum of previous entry rows, which are the number of hidden neurons and their input connections *shared* by contacts σ_{ij} and σ_{kl}. For example, a comparison of NN3 in Table IV (σ_{13}) and Table XXIV (σ_{14}) would show that these two networks share one hidden neuron with input connections from inputs 1 and 6. The net result, tabulated in the last row entry in Tables V–VII, indicates which network architectures perform *optimally* (all reproduce the database perfectly). The total number of network variables (sum of all weights and biases for all contacts) is consistently 50% of the number of contact observables (number of possible *i,j* contacts times the number of sequences) regardless of sequence length. There is quite a large spread around the 50% average when individual contacts are considered, and we return to this point in Sec. IV.

The relative performance of the three networks can be attributed to the interaction potential, the polypeptide length under consideration, and the convention of our retaining sequences with a larger number of A's when scan-

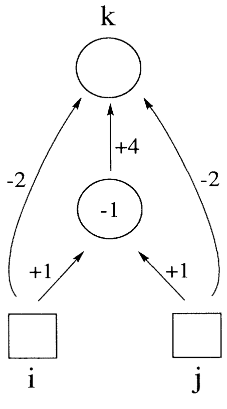

FIG. 10. Central Boolean function for neural network 3.

TABLE IV. Network architectures for heptamer contact σ_{13}. Table rows refer to the entire network architecture when using the central Boolean functions in Fig. 8 (NN1), Fig. 9 (NN2), and Fig. 10 (NN3), respectively. Columns denoted by "Weights" refer to the weight variables found for the optimal network for connections between the input channels and hidden layer neurons ($i \rightarrow h$), hidden neurons to output neurons ($h \rightarrow o$), and input channels connected directly to the output neurons ($i \rightarrow o$). The ordering of weights for the $i \rightarrow h$ and $i \rightarrow o$ columns imply weight subscripts w_{1j}, w_{2j}, w_{3j}, etc., where the numerical values refer to the first, second, third, etc., amino acids of a sequence as scanned from left to right, and j is the hidden or output neuron of interest. The columns denoted by b_h and b_o are the value of the bias for hidden neurons and output neurons, respectively.

σ_{13}	Weights ($i \rightarrow h$)	b_h	Weights ($h \rightarrow o$)	Weights ($i \rightarrow o$)	b_0
NN1	0,0,0,1,0,−1,−1	+2	+1	1,0,1,0,0,0,0	−3
	0,−1,0,0,0,1,−1	+2	+1		
	−1,−1,0,0,−1,0,1	+3	+1		
	0,−1,0,−1,1,−1,0	+3	+1		
	−1,1,−1,−1,0,−1,0	−4	+2		
NN2	1,0,0,−1,0,0,0	+1	+1	−1,0,−1,0,0,0,0	+4
	1,0,0,0,−1,0	+1	+1		
	1,1,0,0,0,0,0	+1	+1		
	1,−1,1,0,0,0,0	−2	+2		
	0,0,0,1,1,1,1	−3	+2		
	1,0,1,0,−1,−1,−1	−4	+2		
	1,0,1,−1,−1,0,−1	−4	+2		
NN3	1,0,1,0,0,0,0	−1	+4	−2,0,−2,0,0,0,0	−7
	1,1,0,0,0,0,0	+1	+1		
	1,0,0,0,0,−1,0	+1	+1		
	1,0,0,−1,0,0,0	+1	+1		
	0,−1,0,1,0,0,−1	+2	+1		
	0,−1,0,0,1,0,−1	+2	+1		
	0,−1,0,0,0,1,−1	+2	+1		
	−1,−1,0,0,−1,0,1	+3	+1		
	0,−1,0,−1,1,−1,0	+3	+1		

ning left to right (see Sec. II). Between these three architectures, NN2 is clearly the least optimal of all networks; this is most certainly due to the interaction potential [Eq. (2.1) and (2.2)] whereby A_iA_j contacts are more highly favored then B_iB_j contacts, so that the Boolean function which characterizes NN2 incorrectly disables the more prevalent positive value contact and overenables the more unlikely B_iB_j interaction. For this reason NN1 performs optimally for the case of the pentamer and hexamer and quite well for the heptamer. However, NN3 performs optimally as the polypeptide length increases to the heptamer length. The convention of retaining a particular sequence direction results in the greater success of NN1 for predicting contacts at the beginning of the se-

TABLE V. Summary of network variables for NN1, NN2, and NN3 for the 2D pentamer. A summary of the number of weights and biases for perfect network architectures for NN1, NN2, and NN3 (see text). The row denoted "shared" indicates the number of hidden neurons and their input connections which are redundant between two or more σ_{ij} contacts. The row denoted "total" provides the total number of network variables required to predict the tertiary structure of all sequences.

Pentamer	NN1		NN2		NN3	
contact	Weights	Biases	Weights	Biases	Weights	Biases
σ_{13}	10	2	16	4	12	2
σ_{14}	8	1	12	3	11	2
σ_{15}	9	2	10	2	10	1
σ_{24}	13	3	15	5	17	4
σ_{25}	10	2	12	3	11	2
σ_{35}	17	3	14	3	16	4
Shared	7	2	14	6	3	1
Total	60	11	65	14	64	14

TABLE VI. Summary of network variables for NN1, NN2, and NN3 for the 2D hexamer. See Table V caption.

Hexamer	NN1		NN2		NN3	
contact	Weights	Biases	Weights	Biases	Weights	Biases
σ_{13}	11	2	19	5	14	4
σ_{14}	17	4	25	6	20	6
σ_{15}	19	4	23	5	18	4
σ_{16}	15	4	15	3	15	3
σ_{24}	25	5	28	6	24	5
σ_{25}	20	4	24	6	21	5
σ_{26}	26	5	24	6	22	6
σ_{35}	23	4	20	5	22	6
σ_{36}	11	2	11	2	12	3
σ_{46}	17	3	14	3	16	4
Shared	35	10	16	7	17	7
Total	149	27	187	40	167	39

quence, such as σ_{13}, because a larger pool of A_1A_3 sequences are present at this i,j position (see Tables I–III). NN2 predicts contacts optimally at the other end of the chain (σ_{35} in the pentamer case, for example) because of the greater number of $B_{n-2}B_n$ sequences for this contact. NN3 becomes the optimal network as the polypeptide length increases sufficiently to minimize end effects, so that the number of A_iA_j and B_iB_j sequence possibilities for interior i,j contact position are equal, or nearly so. NN1 still competes effectively with NN3 for network optimality of a small number of intermediate heptamer contacts due to the greater number of positive A_iA_j contacts relative to the number of positive B_iB_j contacts, due to the interaction potential described in Sec. II. However, when summed over all contacts, NN3 is superior.

The polypeptide lengths which we have considered here are much smaller than those conventionally seen in real protein databases. However, the heptamer length is sufficient for establishing the network topology trend toward large polypeptide sequences, and the Boolean function in Fig. 10 is clearly the optimal of the three networks considered in the limit of large sequence length. We note that a combination of network topologies is optimal for the pentamer and heptamer by strict definition (i.e., a minimum in the number of network variables by choosing NN1, NN2, or NN3 for *each* contact). However, NN3 is competing effectively with the combination network at the heptamer length and will likely become optimal in the length limit due to a greater number of shared hidden neurons between contacts. Furthermore,

TABLE VII. Summary of network variables for NN1, NN2, and NN3 for the 2D heptamer. See Table V caption.

Heptamer	NN1		NN2		NN3	
contact	Weights	Biases	Weights	Biases	Weights	Biases
σ_{13}	26	5	32	7	36	8
σ_{14}	20	4	32	7	23	6
σ_{15}	36	7	44	9	33	8
σ_{16}	41	7	39	8	26	5
σ_{17}	15	4	14	2	17	4
σ_{24}	32	6	41	9	39	9
σ_{25}	44	9	49	9	40	9
σ_{26}	54	10	59	12	51	10
σ_{27}	32	6	39	9	27	6
σ_{35}	36	7	33	7	38	8
σ_{36}	43	8	47	10	39	9
σ_{37}	27	5	32	7	26	7
σ_{46}	34	6	27	6	29	7
σ_{47}	34	6	30	6	27	7
σ_{57}	38	6	30	7	32	8
Shared	34	9	20	7	32	13
Total	478	87	528	108	451	98

the potential function described by Eqs. (2.1) and (2.2) may be somewhat exaggerated in its interaction strength between residues relative to that in real proteins. NN3 should have a greater chance of success as more subtle interactions are introduced into a force field, since it does not discriminate between degrees of attractive interactions, but instead differentiates on commonly observed, and ultimately transferable, physical-chemistry traits of attractive versus repulsive interactions. Based on the above analysis, we believe that the architecture of NN3 is the most robust for transfer to larger polypeptides in two dimensions and ultimately to real protein studies of interest.

IV. RESULTS AND DISCUSSION

In addition to the highly desirable trait of neural-network predictive accuracy, insight into the means by which neural networks learn the higher-order information in sequence-structure correlations, i.e., the chemistry of amino-acid interactions, would be equally valuable. It has been a well-documented fact that "good" training on a database may actually result in overtraining so that generalization by the associative-memory algorithm is lost [1,8,15]. In regard to the protein-folding problem, this loss of generalization results in the inability of the network to recognize chemical relationships between sequence and structure. We believe that the optimal neural networks presented in the preceding section for our two-dimensional polypeptides indicate that insightful learning is possible and may provide clues as to how networks can be devised and trained to predict the mapping between sequence and tertiary structure for more realistic polypeptides and proteins. In this section we comment on the traits that optimal neural networks should satisfy in order to bring to fruition both prediction accuracy and genuine learning strategies. As concluded in the preceding section, we believe that NN3 (Fig. 10) is the best general architecture for neutral-network predictions of protein structure, and it is on this network which we base the following discussion.

The central Boolean function depicted in Fig. 10 alone recognizes a very important relation between sequence and structure. The potential-energy function defined in Eqs. (2.1) and (2.2) strongly favors interactions between A_iA_j residues, and to a lesser extent, B_iB_j residues, while A_iB_j (B_iA_j) interactions are unfavorable. The recognition of this particular sequence-structure relationship results in 74–78 % contact prediction accuracy for NN3 before correction. However, the *representation* of the input is crucial for exploiting the architectural feature just described. Our input representation of amino acids A and B of $+1$ and -1, respectively, would be indicative of a chemical feature of self-attraction, such as interaction between two hydrophobic groups, which naturally arises in the interaction potential described in Eqs. (2.1) and (2.2). Given the appropriate input representation, the architecture of NN3 exploits this relevant "second"-order information [13] by succinctly encoding the nonbonded interaction into the central Boolean function.

In addition, long-range sequence information is required to predict even local contacts. The dedicated neuron indicated by $1,0,0,0,0,-1,0$ input weights for the σ_{13} contact for the heptamer in Table IV provides such an example, where an amino acid far removed from the pair under consideration dictates quite strongly the fold outcome. Recent neural-network applications which use "windows" of amino acids, i.e., local sequence information, for predicting distance matrices [19,20] of realistic polypeptides and protein structures would be deficient in two respects. First, prediction accuracy is lost to some significant extent. We have found that *many* corrections to the central Boolean function are possible with only one hidden neuron with input from at least this one important amino acid; this is due to common sequence information among a number of sequences which signals a folding outcome. Second, insight may be lost as well. It is quite plausible that amino acid(s) outside the window may be an especially important residue marker for signaling a particular fold, as in the σ_{13} example exhibited by our two-dimensional (2D) polypeptides described above. In a related vein, shared hidden neurons between contact outputs may have special significance: particular amino acids (those whose weight connections to that neuron are nonzero) may be lynchpin residues which determine compound structural features. For example, in the case illustrated by the heptamer contact σ_{13}, amino acid 7 determines whether contacts 13, 14, and 15 are on or off, as well as simultaneous signaling different spatial regions (contacts 13 and 46) of the native state. The propensity of shared neurons for NN3 increased with polypeptide size.

We have also found that a small change in sequence can result in significant changes in fold (in our model as to whether a contact is on or off) and that networks do indeed have more trouble with this case. We find that a dedicated hidden neuron with virtually full input connectivity is required in this case to correctly predict the fold outcome. We have found that elimination of problem sequences (the necessity of having one dedicated neuron to predict one contact of one sequence correctly) does not degrade overall performance significantly. For NN3 the total number of incorrectly predicted sequences is 5 out of 120 for the pentamer, 13 out of 360 for the hexamer, and 31 out of 1080 for the hexamer. Thus the acceptance of some performance degradation such as this may be desirable in a pragmatic sense, where we find a 33% reduction in the number of hidden neurons and the number of weights with only 3–4 % performance loss.

Another more subtle aspect of the interplay between network design and genuine learning is the number of hidden layers which optimally (fewest number of network variables) and accurately predict all contact values for all sequences. Although we have explored network designs with two hidden layers, never did we find such an architecture which was more optimal than those presented in Tables IV–XXX, i.e., architectures containing only one hidden layer. This seems to be consistent with the network complexity required by our model with only two amino-acid types; only with greater sequence diversity should additional hidden layers be important. By the same token, we rarely found cases where corrections to the central Boolean function involved direct connectivity from the input to the output, indicating that our 2D polypeptide model is not grossly oversimplified.

The networks devised for tertiary structure prediction of 2D polypeptides may also provide some guidance in overcoming the well-appreciated multiple-minima problem in the space of the network variables [1]. In general, backpropagation algorithms are preferred in spite of the fact that only local minimum solutions are found. While researchers may train networks several times with differing initial guesses to address this deficiency of backpropagation optimization, the initial neural-network topologies are always fully connected (all weight and bias variables started with nonzero values), and network variables which become zero during optimization are not "pruned out" as sometimes suggested [1]. Our optimal networks (close to a network global solution) indicate that sparse input-hidden neutron connections (some weights equal to zero) are often successful at predicting the contact outcome correctly for many sequences, while dense input-hidden neuron connectivity only predicted the contact correctly for one, or a very small number, of sequences. These network-design results indicate that weights which are initially zero, or become so during backpropagation, should remain so in order to avoid unprofitable regions of network variable space.

In summary, the construction of neural networks for our 2D polypeptides with two amino-acid types indicates that genuine learning strategies are present in the networks. It is evident from the above observations that the central Boolean function described in Fig. 10 successfully maps a basic sequence-sequence structure relationship—namely the nonbonded interaction potential in Eqs. (2.1) and (2.2). When the central Boolean function (or nonbonded interaction) is not sufficient for accurately describing the contact value, additional hidden neurons are then required to understand the interplay between the nonbonded interactions *and* connectivity portions of the potential. In these cases, the additional hidden neurons may provide insight into the more general aspects of the protein-folding problem. The degree of network connectivity necessary between the input and these additional hidden neurons and how many sequences are affected by these dedicated neurons provide a means for refining the protein folding problem as one of strong or weak sequence-structure mappings [27]. The networks presented here for our simple 2D model have demonstrated that very few contacts rely on strong sequence-structure relationships, as indicated by the small degradation in performance when dedicated neurons which correctly predict one contact value for one sequence are eliminated. Hidden neurons with sparse input-hidden neuron connectivity which aid in the correct contact value prediction for many sequences imply that good network designs might be able to reveal important residues which dictate the folding outcome. Finally, thoughtful neural-network design may deconvolute the problem of multiple minima in the space of the network variables by isolating the relevant region with informed initial guesses for weights and biases.

V. CONCLUSIONS AND FUTURE DIRECTIONS

The successful application of neural networks for a particular pattern recognition task is stymied by three problem areas: the poor quality of the database on which the networks train, network architectures and input representations which are improperly chosen, and the mathematical optimization difficulties faced in finding the best solution in a multiple-minima network solution space. The prediction of protein sequence-structure correlations provides a perfect example of a neural-network application which incorporates these three detrimental traits. In this work we have chosen to take a step back from more ambitious attempts [5–20] at protein-structure prediction by defining a simplified polypeptide model residing in two dimensions with only two amino-acid types. This allows us the advantage of determining the native structure for all possible sequences in order to define a perfect and complete database, at least for polypeptides ranging up to heptamer lengths. With this complete database we have determined network architectures which both accurately and, we believe, optimally (fewest number of network variables) reproduce the observed database of sequence-structure relationships. In this remaining section, we discuss how the insight gained in this model study might impact on neural-network predictions of proteins in three dimensions with full sequence diversity.

One of the most positive conclusions to be derived from this work is that genuine neural network insight into sequence-structure mappings is possible when architectures are thoughtfully designed. In fact, the nonbonded interaction mathematically described in Eqs. (2.1) and (2.2) has been encoded into the central Boolean function (Fig. 10) of our most robust network design of our 2D model chemistry. We emphasize that the representation of the sequence input is particularly important for exploiting this design feature fully. When transferring this particular architecture to the full complexity of real proteins, an ordering of the 20 amino acids on a hydrophobic scale [19] ranging from +1 (most hydrophobic) to −1 (least hydrophobic) may be useful. In this case, the central Boolean architecture will at least distinguish between the unlikely hydrophobic-polar contacts from the polar-polar and hydrophobic-hydrophobic contacts, so that other hidden neurons can address the subtleties of the relative nature of the attractive nonbonded interactions such as hydrogen bonding and salt-bridge formation. If a different representation is chosen where interactions are repulsive (net electrostatic monopole assignments of the amino acids, for example), then the topology of the Boolean function would still be used, but with the weight assignments depicted as in Fig. 11.

The suggestion [13,14] that full exploitation of hidden-layer architectures is not possible due to a scarcity of examples in a sparse protein database may be unduly pessimistic until alternative input representations and architectures are thoroughly explored. Our simple chemical model required a neural network with one hidden layer to successfully predict its structure. Recent predictions of secondary structure using neural networks have shown

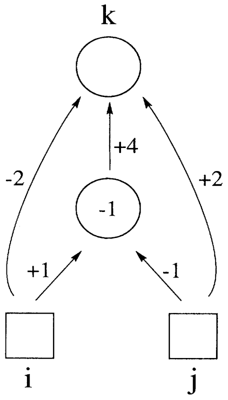

FIG. 11. Central Boolean function for neural network 3, which best exploits an "opposites attract" potential-energy interaction.

[8] that hidden-layer architectures perform better than architectures with no hidden layers and appear to be abstracting higher-order correlations which can be understood from a chemical point of view. Demonstrating that the central Boolean function discussed above actually has some impact on improving neural-network tertiary-structure prediction would also be necessary to further enhance this point.

In our model, further corrections to the central-Boolean-function prediction are required when the backbone bend portion of the potential plays as an important role as the nonbonded interaction in determining the native structure. The model suggests that additional hidden neurons which provide these corrections may also provide important insight as well. When many corrections are possible with one hidden neuron, then the amino-acid positions connected to this neuron may be an important signal for the fold outcome, analogous to known cases where certain amino acids signal initiation or termination of secondary structure [28] for example. Whether new sequence-structure signals can be ultimately uncovered by neural networks in protein structure prediction remains an exciting possibility.

The self-contained model presented here permits several obvious future pursuits in the area of protein structure prediction by neural networks. First, we can systematically investigate the impact of database degradation on the design of optimal neural-network predictions. This is a particularly insidious problem in the protein-folding area since virtually all prediction methods [5–26,29,30] rely to some or significant extent on learning examples; it is painfully clear that homology modeling, statistical methods, and neural-network applications perform best when applied to protein structural classes [10,29,30] or to structures with strong sequence homologies [19,20]. The polypeptide model and neural-network model architectures described here should allow a systematic exploration of what degree of prediction accuracy is possible when particular structural classes or sequence homologies are exploited in the database and to what degree either representation is useful. The effect of finite polypeptide length input "windows" can also be investigated with our simple model database of pentamers, hexamers, and heptamers.

The neural-network solutions for our simplified protein model also provide fertile ground in which mathematical optimization techniques can be devised for determining optimal neural-network solutions [31,32]. One such optimization problem is the slow convergence exhibited by the backpropagation algorithms for neural-network learning; the other, and more difficult, problem is the multiple solutions available in the space of the network variables given a network architecture. While it is clear that steepest-descent algorithms are inferior for quick convergence to a local minimum, what particular second-order methods are optimal can be explored with our simplified polypeptide neural-network model. Similarly, simulated annealing methods for addressing the global minimization problem rely on the determination of an optimized cooling schedule for converging to the global minimum; the search for such a cooling schedule should be more feasible with our stripped-down model. In both cases we believe our model shows sufficient complexity so that optimization methods developed for this case are transferable to a more realistic 3D model. In relation to this last point, the network topologies derived from our 2D model may also overcome the deficiencies of backpropagation training of real proteins, by providing good initial guesses of the network variables that converge to a local minimum that is "near" the global solution.

Ultimately, the goal of this continued neural-network study is to apply the principles learned for this simplified version of neural-network predictions of 2D structures to the genuine prediction of tertiary structure in three dimensions with full sequence diversity, given the accompanying limitations of the database. The accurate determination of even the residue-residue contact values in this case would provide a reasonably robust protein-structure prediction method. While we believe that further important enhancements to straight network structure predictions are still feasible (and necessary), it is unlikely that complete and accurate structure prediction by neural networks alone is possible. Nonetheless, we are prepared for this outcome with an optimization method [5] which incorporates neural-network predictions into empirical protein force fields as guidance for smoothing the complex protein hypersurface to retain only the native-structure minimum. Frustrated interactions resulting from the interplay of the protein potential-energy function and constraints representing the neural-network predictions serve to aid the search for an optimal structure determined with full atomic resolution. A pilot study of the method applied to melittin [5] showed that the deficiencies of neural network predictions can be redressed by their incorporation into an empirical protein force field to provide predicted structure in excellent agreement with the crystal structure. To what extent degradation of neural-network performance is permissible for robust protein tertiary-structure prediction of globular proteins is the topic of future studies in the further development of constrained optimization.

[1] B. Müller and J. Reinhardt, *Neural Networks: An Introduction* (Springer-Verlag, Berlin, 1990).

[2] C. C. Klimasauskas, IEEE Commun. Mag. **30**, 50 (1992).

[3] A. Troll and W. Feiten, Cybern. Syst. **23**, 447 (1992).

[4] K. Y. Tam and M. Y. Kiang, Manage. Sci. **38**, 926 (1992).

[5] T. Head-Gordon and F. H. Stillinger, Biopolymers **33**, 293 (1993).

[6] P. Stolorz, A. Lapedes, and Y. Xia, J. Mol. Biol. **225**, 363 (1992).

[7] S. Hayward and J. F. Collins, Proteins Struc. Funct. Genetics **14**, 372 (1992).

[8] S. M. Muskal and S. H. Kim, J. Mol. Biol. **225**, 713 (1992).

[9] M. Vieth and A. Kolinski, Acta Biochim. Pol. **38**, 335 (1991).

[10] D. G. Kneller, F. E. Cohen, and R. Langridge, J. Mol. Biol. **214**, 171 (1990).

[11] M. J. McGregor, T. P. Flores, and M. J. E. Sternberg, Protein Eng. **2**, 521 (1989).

[12] L. H. Holley and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **86**, 152 (1989).

[13] N. Qian and T. J. Sejnowski, J. Mol. Biol. **202**, 865 (1988).

[14] M. J. Rooman and S. J. Wodak, Nature **335**, 45 (1988).

[15] J. D. Hirst and M. J. E. Sternberg, Biochemistry **31**, 7211 (1992).

[16] E. A. Ferran and P. Ferrara, Comput. Appl. Biosci. **8**, 39 (1992).

[17] J. D. Hirst and M. J. E. Sternberg, Protein Eng. **4**, 615 (1991).

[18] Y. Bengio and Y. Pouliot, Comput. Appl. Biosci. **6**, 319 (1990).

[19] H. Bohr, J. Bohr, S. Brunak, and R. M. J. Cotterill, FEBS Lett. **261**, 43 (1990).

[20] G. L. Wilcox, M. Poliac, and M. N. Liebman, Tetrahedron Comput. Methodol. **3**, 191 (1990).

[21] J. M. Levin, B. Robson, and J. Garnier, FEBS Lett. **205**, 303 (1986).

[22] J. Garnier, D. J. Osguthorpe, and B. Robson, J. Mol. Biol. **120**, 97 (1978).

[23] P. Y. Chou and G. D. Fasman, Biochemistry **13**, 222 (1974).

[24] V. I. Lim, J. Mol. Biol. **88**, 873 (1974).

[25] O. B. Ptitsyn and A. V. Finkelstein, Protein Eng. **2**, 443 (1989).

[26] J. F. Gibrat, J. Garnier, and B. Robson, J. Mol. Biol. **198**, 425 (1987).

[27] T. Head-Gordon, F. H. Stillinger, M. H. Wright, and D. M. Gay, Proc. Natl. Acad. Sci. U.S.A. **89**, 11 513 (1993).

[28] L. G. Presta and G. D. Rose, in *Protein Folding: Deciphering the Second Half of the Genetic Code,* edited by L. M. Gierasch and J. King (American Association for the Advancement of Science, Washington, D.C., 1990), p. 29.

[29] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918 (1992).

[30] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 9029 (1992).

[31] J. A. Kinsella, Network Comput. Neural Syst. **3**, 27 (1992).

[32] E. Barnard, IEEE Trans. Neural Networks **3**, 232 (1992).

[33] See AIP document no. PAPS PLEEE8-48-1502-21 for 21 pages of tables giving optimal network solutions. Order by PAPS number and journal reference from American Institute of Physics, Physics Auxiliary Publication Service, 335 East 45th Street, New York, NY 10017. The price is $1.50 for each microfiche (60 pages) or $5.00 for photocopies of up to 30 pages, and $0.15 for each additional page over 30 pages. Airmail additional. Make checks payable to the American Institute of Physics.