Sequence design in lattice models by graph theoretical methods

B. S. Sanjeev, S. M. Patra*, and S. Vishveshwara^
Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India

A general strategy has been developed based on graph theoretical methods, for finding amino acid sequences that take up a desired conformation as the native state. This problem of inverse design has been addressed by assigning topological indices for the monomer sites (vertices) of the polymer on a $3\times3\times3$ cubic lattice. This is a simple design strategy, which takes into account only the topology of the target protein and identifies the best sequence for a given composition. The procedure allows the design of a good sequence for a target native state by assigning weights for the vertices on a lattice site in a given conformation. It is seen across a variety of conformations that the predicted sequences perform well both in sequence and in conformation space, in identifying the target conformation as native state for a fixed composition of amino acids. Although the method is tested in the framework of the *HP* model [K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989)] it can be used in any context if proper potential functions are available, since the procedure derives unique weights for all the sites (vertices, nodes) of the polymer chain of a chosen conformation (graph).

I. INTRODUCTION

The problem of sequence design has dominated the field of protein folding for about two decades.[1–14] Innumerable efforts have been made experimentally and theoretically to successfully design sequences of amino acids that fold rapidly into their desired target native state.[1,14–19] Computationally lattice models are amenable to exact enumeration of all possible compact conformations of a polymer. Therefore lattice models are a testing ground for ideas that could be useful in the design of real proteins. The *HP* model introduced by Lau and Dill[20] is perhaps the simplest model that captures some of the essential features of real proteins such as hydrophobicity and compactness. In general, the probability that a sequence S is housed in a conformation Γ is given by

$$P(S,\Gamma)=\frac{\exp[-H(S,\Gamma)/kT]}{\exp[-F(S)/kT)]}, \tag{1.1}$$

where $H(S,\Gamma)$ is the energy of the sequence in conformation Γ, T is the temperature, and the sequence free energy $F(S)$ is defined by the equation

$$\exp(-F(S)/kT)=\Sigma_{\Gamma'}\exp[-H(S,\Gamma')/kT]. \tag{1.2}$$

In the simple case that F is sequence independent for a target structure Γ, $P(S,\Gamma)$ is maximized on determining a sequence S^* for which $H(S^*,\Gamma)$ is as large and negative as possible.[21] It has been argued that $F(S)$ is substantially sequence independent as long as the composition is fixed.[22]

It is recognized that the topology of the target conformation is a key feature in the process of protein folding.[23–27] Our present design strategy is simple and focuses only on the topology of the target conformation. We have developed topological indices which capture the features of the native conformation by graph theoretical methods. The method is highly successful as the designed sequences, in a fixed composition space, perform well for thirteen native structures of diverse topology on a $3\times3\times3$ cubic lattice model.

Graph theory in general is an important tool in topological investigations. Chemical graph theory has been used to investigate the topological properties of covalently bonded systems.[28–30] In the context of proteins, this technique has been used in the identification of secondary structures,[31] recognition of beta-folds in beta barrels,[32] and tertiary folds in *G*-proteins[33] from sequence information. Vibrational dynamics and thermal fluctuations in proteins have also been characterized by graph theoretical methods.[34–37] We have been interested recently in representing the noncovalent interactions in proteins[38] and recognizing the main-chain[23] and the side-chain clusters and cluster-centers[24] by graph theoretical methods. From these investigations, we had demonstrated that the highest eigenvalue and their vector components of a connectivity matrix of a protein carry topological information of the graph. Using this method, the structural features such as the conserved clusters in a family of proteins, nucleation sites, active-site clusters have been characterized.

In the present study, we have further explored the topological features and have shown that they can be effectively used for the purpose of inverse folding. Given a conformation [graph] of a polymer, we have been successful in arriving at unique weights for different sites [vertices] of the polymer in the conformation and with these weights, sequences that take up the desired conformation as their native structures have been designed. The design strategy in the present study has been adopted in the *HP* framework on a cubic lattice. However the concepts can be easily extended to bigger lattices and possibly to real proteins. Since unique weight on each vertex of the graph is generated, the method can also go beyond *HP* model if appropriate potential func-

*Present address: Polymer Engineer, PSM, John F. Welch Technology Centre, Bangalore 560 066, India.
^Author to whom correspondence should be addressed. Telephone: +91-80-3092611; Fax: +91-80-3600535. Electronic mail: sv@mbu.iisc.ernet.in

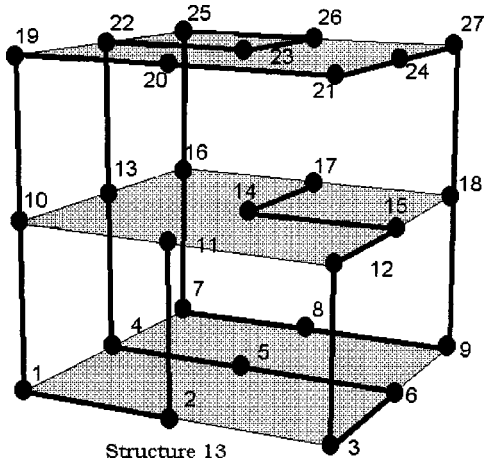

FIG. 1. A 3×3×3 lattice with the lattice points numbered. The structure of a 27 residue polymer [Structure 13, Table I] is represented by a solid line connecting the lattice points.

tions are available. There are two major parts of designing sequences for a target structure. First is that these sequences should have the lowest energy amongst all possible sequences (sequence space). More importantly, the predicted sequences should identify the target structure as their native structure amongst all possible structures (structure space). In the context of real proteins, both these aspects are highly challenging, considering the sizes of sequence and structure space. Our present methodology focuses on identifying best sequences in the sequence space for a given composition. It is seen that sequences that are very good in the sequence space are also well behaved in the structure space in terms of identifying the chosen structure as the native. The methodology developed is described in the next section and the evaluation of the performance of the designed sequences is presented in Sec. III.

II. METHODS

A. Structures on three-dimensional lattice

Lattice models are extremely useful in enumerating all compact structures of a polymer and evaluating the noncovalent interaction energies. In the present study, we have considered 27 residue polymer structures on a 3×3×3 cubic lattice. It is known that 103 346 independent compact self-avoiding [CSA] structures can be generated on a cubic lattice.[39,40] An inspection of cubic lattice (Fig. 1) shows that there are four types of vertices—(a) one body center vertex (14), (b) six face centered vertices (5, 11, 13, 15, 17, 23), (c) twelve edge center vertices (2, 4, 6, 8, 10, 12, 16, 18, 20, 22, 24, 26), and (d) eight corners (1, 3, 7, 9, 19, 21, 25, 27). These four types of vertices when occupied by nonterminal residues can have respectively 4, 3, 2, and 1 noncovalent neighbors in the lattice (nonbonded connectivity of the vertices) and have 5,4,3, and 2 neighbors otherwise. We have classified all the 103 346 CSA structures into three categories based on the distribution of twenty-eight nonbonded connectivities (edges) between different types of vertices. We have denoted them as classes $A-C$ in which there are one, two and three vertices of highest connectivity (4) [degree], respectively. The number of vertices with degrees 4, 3, 2, and 1 are 1, 6, 14, and 6, respectively, in class *A*, which is denoted as 4(1)-3(6)-2(14)-1(6). Similarly, the distribution of vertices with different degrees in classes *B* and *C* are 4(2)-3(5)-2(13)-1(7) and 4(3)-3(4)-2(12)-1(8), respectively. There are 40 144, 50 588, and 12 614 number of CSA, respectively, in classes *A*, *B*, and *C*. A similar classification of CSA structures of a 18 residue polymer on a 3×2×2 lattice was earlier done by Demirel *et al.*[36]

A set of thirteen structures, given in Table I, was chosen for design study. Our aim was to select a range of CSA structures which are unbiased representatives of different classes. About six hundred structures were identified as designable based on the fact that at least one *HP* sequence with KGS potential[41] takes these structures as unique ground state. We separated these structures into classes *A*, *B*, and *C* and obtained eigenvalues of the adjacency matrices of these structures [method described below]. The eigenvalues of each structure were sorted and the highest eigenvalues [Hev] of all the structures were rank ordered. Two structures with high Hev [+] and two structures with low Hev [−] in each of the three classes were selected for further investigations.

TABLE I. Selected structures built on the lattice points given in Fig. 1.

SNo[a]	Cl[b]	Structure[c]																									
1	$A-$	1	2	5	4	7	16	25	22	23	24	21	20	11	12	3	6	9	8	17	26	27	18	15	14	13	10	9
2	$A-$	1	2	3	6	15	12	11	14	5	4	7	8	9	18	17	26	27	24	23	22	25	16	13	10	19	20	21
3	$A+$	1	2	5	14	23	24	15	6	3	12	21	20	11	10	19	22	13	4	7	16	25	26	17	8	9	18	27
4	$A+$	1	2	5	14	11	10	13	4	7	16	17	8	9	18	27	26	25	22	19	20	23	24	21	12	15	6	3
5	$B-$	5	14	11	2	1	10	19	22	13	4	7	8	9	6	3	12	15	24	27	18	17	16	25	26	23	20	21
6	$B-$	1	2	5	8	9	6	3	12	15	24	21	20	11	10	19	22	25	26	27	18	17	16	7	4	13	14	23
7	$B+$	1	2	3	6	5	4	7	8	9	18	27	26	25	22	23	24	21	20	19	10	11	12	15	14	13	16	17
8	$B+$	1	2	3	6	5	4	7	8	9	18	17	16	25	26	27	24	23	22	13	14	15	12	21	20	19	10	11
9	$C-$	5	14	11	10	19	20	21	24	27	18	17	8	9	6	15	12	3	2	1	4	7	16	13	22	25	26	23
10	$C-$	5	2	1	10	11	12	3	6	9	18	27	26	23	14	15	24	21	20	19	22	25	16	13	4	7	8	17
11	$C+$	5	2	1	10	19	20	11	14	13	4	7	8	9	6	3	12	21	24	27	26	23	22	25	16	17	18	15
12	$C+$	5	2	11	14	13	16	17	8	7	4	1	10	19	20	23	22	25	26	27	24	21	12	3	6	9	18	15
13[d]	C	11	2	1	10	19	20	21	24	27	18	9	8	7	16	25	26	23	22	13	4	5	6	3	12	15	14	17

[a]Structure number.
[b]Class ($A/B/C$); +/− stand for higher/lower Hev conformations.
[c]Structures are represented by vertex numbers (27 residues), taken sequentially by the polymer chain.
[d]Highly designable structure (Ref. 42), also shown in Fig. 1.

Further, the unique structure, which was identified as highly designable by Li *et al.*,[42] was also selected.

B. Graph representation

Graphs can be represented in algebraic form as matrices and molecular graphs are generally constructed to represent chemical structures.[43,44] Recently we have used graph theoretical representation to indicate nonbonded interactions in polymers on lattice points[38] and in real protein structures[23,24] and have shown that structural classification can be made by this method. We have also shown that the eigenvalues of adjacency or Laplacian matrix can be correlated with the nature of branching of the graph and that the corresponding vector components carry information on the position of nodes in the graph.[24] The structural characterization and the present method of designing sequences for a given structure are based on the eigenvalues and eigenvectors of Laplacian matrix, which is defined as $L=D-A$, where A is the adjacency matrix and D is the degree matrix.[24]

The adjacency matrix ($A=[a_{ij}]$) is a square symmetric matrix of the order of the number of monomer units/lattice sites, which is 27 for a $3\times3\times3$ lattice. The elements of this matrix, $\{a_{ij}\}$, are defined as:

$a_{ij}=1$ if ith and jth monomers occupy adjacent lattice points

$=0$ otherwise, the elements a_{ik} are also set to zero when $k=i-2$ to $i+2$.

Diagonal matrix (D) is a matrix of the same order as the adjacency matrix with diagonal element at the ith row/column element being the number of nonbonded interactions of the ith center. The Laplacian matrix is diagonalized to obtain eigenvalues and eigenvectors.

C. Sequence design

1. Ranking the vertices by graph theoretical parameters

In this study we have further investigated the eigenvalues and eigenvectors in order to obtain the topological index for every vertex. Such an index can be used as a weight that determines the nature of the vertex, for example, hydrophobic or polar. Each vertex has a unique place on the graph and the developed indices capture its place in the graph. The details of the development of such indices are described here.

There are two important parameters, which determine the weight of a vertex on a graph—the degree of the vertex, and the position of the vertex in the global context of the graph. We have earlier shown[24] that the eigenvector components $[Hvc]$ of the highest eigenvalue [Hev] carry the information about the position of the vertices in the graph. Although, generally it includes the connectivity information such as the vertices of high degree having higher Hvc, it is not necessarily retained if one looks at Hevs alone.[23,24] Hence, the vertex weight is deduced by adding the degree of the vertex to the Hvc of that vertex. This ensures to a great extent that vertices of higher degree are given higher weight. Thus, the vertices of the same degree are discriminated by their position in the graph. In the context of the compact structure of a polymer, the weights discriminate between the core and the surface residues of the same degree since the core residues have connections with vertices of higher degree than the surface residues. This method of weight assignment for vertices works well for structures represented by a single connected graph [Fig. 2(a)].

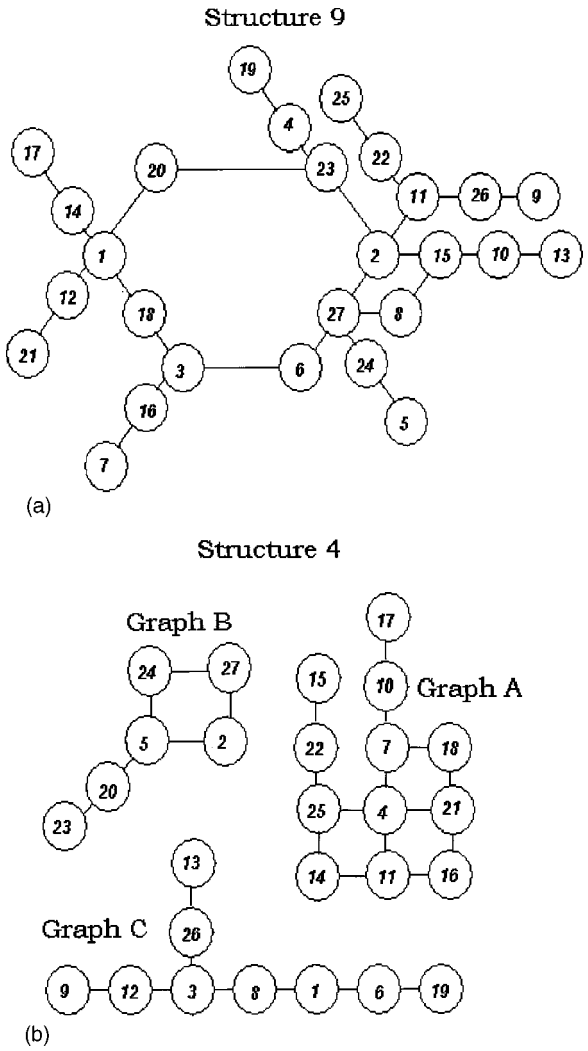

FIG. 2. Two-dimensional graph representation of a conformation represented by (a) a single graph, and (b) three graphs-*A*, *B*, and *C*.

The situation becomes more complicated when the structure is represented by disconnected graphs [Fig. 2(b)], since each graph has its own Hev (the vector components of other graphs being zero, see Table II). The Hev for each of the disconnected graph of a structure is identified and the corresponding $Hvcs$ are chosen as the weights of the vertices making up that graph. In these cases however, not only the positions within a graph have to be weighted, but also they have to be weighted in correlation across graphs. We have considered two scaling factors for ranking the vertices when the structure is represented by more than one disconnected graph. The size of the graph is taken into account by scaling the Hvc by the fraction of (27) vertices constituting the graph. This however is not sufficient because the vertex weight should also take into account the nature of the graph itself.[23,44] This is accomplished by scaling the vector components also by Hev of the concerned graph. The weight evalu-

TABLE II. Calculation of vertex weights and sequence prediction for $10H$-$17P$ composition for Structure 4.[a]

Residue number	Vertex number[b]	Vector components of subgraphs			W_E[c]	R_e	S_e	W_C	$W=W_C+W_E$	R_d	S_d	R_a	S_a
		Graph A	Graph B	Graph C									
1	1	0	0	0.193 22	0.286 24	18	*P*	2	2.286 24	18	*P*	11	*P*
2	2	0	0.394 10	0	0.399 49	14	*P*	2	2.399 49	15	*P*	14	*P*
3	5	0	0	0.741 85	1.098 98	2	*H*	3	4.098 98	2	*H*	7	*H*
4	14	0.678 34	0	0	1.793 80	1	*H*	4	5.793 80	1	*H*	1	*H*
5	11	0	0.701 81	0	0.711 41	7	*H*	3	3.711 41	7	*H*	2	*H*
6	10	0	0	0.089 71	0.132 90	23	*P*	2	2.132 90	21	*P*	8	*H*
7	13	0.315 23	0	0	0.833 59	5	*H*	3	3.833 59	6	*H*	4	*H*
8	4	0	0	0.382 56	0.566 73	8	*H*	2	2.566 73	8	*H*	21	*P*
9	7	0	0	0.100 00	0.148 14	21	*P*	1	1.148 14	23	*P*	26	*P*
10	16	0.084 11	0	0	0.222 42	20	*P*	2	2.222 42	20	*P*	9	*H*
11	17	0.346 13	0	0	0.915 31	3	*H*	3	3.915 31	3	*H*	5	*H*
12	8	0	0	0.344 42	0.510 23	10	*H*	2	2.510 23	10	*H*	19	*P*
13	9	0	0	0.100 00	0.148 14	22	*P*	1	1.148 14	22	*P*	27	*P*
14	18	0.167 43	0	0	0.442 75	12	*P*	2	2.442 75	12	*P*	17	*P*
15	27	0.016 99	0	0	0.044 93	26	*P*	1	1.044 93	25	*P*	24	*P*
16	26	0.175 26	0	0	0.463 46	11	*P*	2	2.463 46	11	*P*	18	*P*
17	25	0.016 99	0	0	0.044 93	25	*P*	1	1.044 93	26	*P*	23	*P*
18	22	0.167 43	0	0	0.442 75	13	*P*	2	2.442 75	13	*P*	16	*P*
19	19	0	0	0.026 05	0.038 59	27	*P*	1	1.038 59	27	*P*	22	*P*
20	20	0	0.307 71	0	0.311 92	17	*P*	2	2.311 92	16	*P*	12	*P*
21	23	0.346 13	0	0	0.915 31	4	*H*	3	3.915 31	4	*H*	6	*H*
22	24	0.084 11	0	0	0.222 42	19	*P*	2	2.222 42	19	*P*	10	*H*
23	21	0	0.086 40	0	0.087 58	24	*P*	1	1.087 58	24	*P*	25	*P*
24	12	0	0.394 10	0	0.399 49	15	*P*	2	2.399 49	14	*P*	15	*P*
25	15	0.315 23	0	0	0.833 59	6	*H*	3	3.833 59	5	*H*	3	*H*
26	6	0	0	0.344 42	0.510 23	9	*H*	2	2.510 23	9	*H*	20	*P*
27	3	0	0.307 71	0	0.311 92	16	*P*	2	2.311 92	17	*P*	13	*P*

[a]S_e, S_d, and S_a are designed sequences (also in Table III); R_e, R_d, and R_a are evaluated ranks of the vertices
[b]See Fig. 1 and Table I.
[c]$Hevs$ for graphs *A*, *B*, and *C* are 5.949 90, 4.561 15, and 4.444 23.

ation of vertices is quantitatively represented as:

$$W_i = W_{iC} + W_{iE}, \tag{2.1}$$

$$W_{iE} = E_i \times f_i \times Hvc(i), \tag{2.2}$$

where W_{iC} is the connectivity of the *i*th vertex, f_i is the ratio of the number of vertices in the graph to which vertex ''*i*'' belongs to the total number of vertices (27) in the structure, E_i is the Hev of the concerned graph and $Hvc(i)$ is the vector component of the vertex ''*i*.''

An example of the scaling and subsequent weight assignment for a structure [Structure 4 in Table I and Fig. 2(b)] is given in Table II. As this structure is made up of three disconnected graphs (*A*, *B*, and *C*), three Hevs corresponding to these graphs are selected. Column 4, for example, shows that graph *B* is made up of six vertices 2, 5, 20, 23, 24, and 27 with an Hev of 4.561 15 and the vertex 5 has the largest Hvc. Similarly, columns 3 and 5 give the information on graphs *A* and *C*. The weights on vertices [*W* in column 10] are derived from the expressions (2.1) and (2.2), based on which the vertex rank [R_d in column 11] is obtained. Ranking gives a unique value for each vertex. Using the ranks (i.e., weights), composition of residues (*H* or *P*) and selecting potential, the type of monomer at every vertex is selected. In our present study, we have used *HP* model with the interaction potential of $H-H=-2.3$, $H-P=-1$ and $P-P=0$.[41] When we confine the composition to $10H$ and $17P$, the first 10 rank ordered vertices take up the *H* type of monomer and the rest of the vertices take up the *P* type monomer in our designed sequence [S_d column 12]. Apart from S_d, sequences S_e [column 8] and S_a [column 14] have also been designed based primarily on vector component [W_E, column 6] and on the degree of the vertex [W_c, column 9], respectively. These designs were done to see if the information from connectivity alone or eigenvector components alone is sufficient or not. Sequences (S_e) were designed by explicitly excluding the connectivity information (W_{iE}) in Eq. (2.1) in order to see how much of information is encoded in Hvc alone. In the other design (S_a), the degree information was retained, but amongst the vertices with equal connectivity the ranking is done such that the vertices with higher weights from eigenvector components have lower ranks. This reverse rank ordering retains the connectivity information of the vertices, however destroys the information on the position of the vertex in the global context of the graph.

2. Sequence generation in HP model

For a given composition the number of different *HP* sequences which can be generated (N_S) is given by

$$N_S = \frac{(n_H + n_P)!}{n_H! n_P!}, \tag{2.3}$$

where n_H in number of *H* and n_P is number of *P* type of monomers.

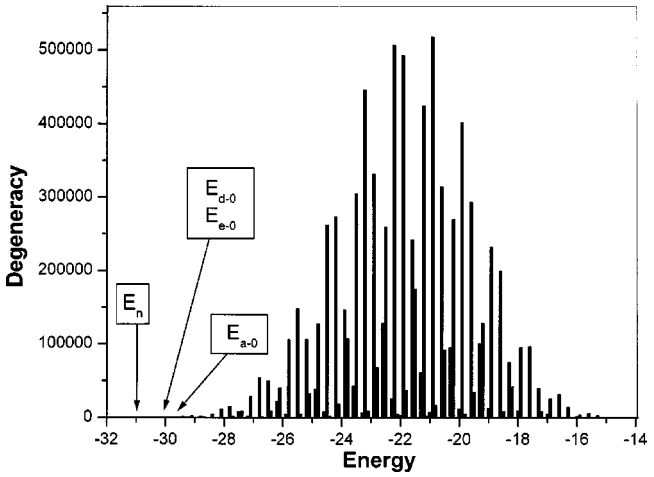

FIG. 3. A typical energy spectrum and degeneracies in sequence space [Structure 4] for a fixed composition [$10H$-$17P$]. The energies of the native sequence S_n and the designed sequences S_d, S_a, and S_e are indicated in the figure successively as E_n, E_{d-0}, E_{a-0}, and E_{e-0}.

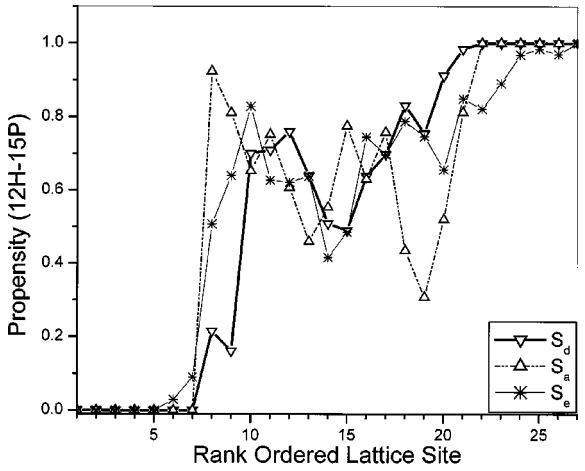

FIG. 4. Propensity plots generated from the structures given in Table I for the composition $12H$-$15P$. The three plots correspond to the rank ordered vertices according to the designed sequences S_d, S_a, and S_e.

The number of sequences is relatively small for a highly biased composition of one type of monomer compared to near equal composition of both types of monomers. All possible sequences were generated for different compositions and the sequences which yield the lowest energy for selected compositions were identified for the structures given in Table I. These sequences denoted as the native sequences (S_n) and they are the best sequences for chosen compositions in the sequence space for the target structure. A typical distribution of sequences and their energy values is given in Fig. 3 for structure 4 and for a composition of $10H$-$17P$. The energy of the native sequence(s) E_n as well as the energies of the designed sequences E_{d-0}, E_{a-0}, and E_{e-0} for this chosen structure are also represented in Fig. 3.

3. Performance evaluation

a. Eluding sequences and energy spectrum in structure space. For a given composition, eluding sequences are defined as those that have lower energies than the designed sequence for a target structure. The design success in sequence space can be evaluated by obtaining the percentage of eluding sequences, success rate being higher for lower percentage of eluding sequences. However, it is more important to evaluate the performance in the structure space since a good designed sequence, designed for its target structure, should have the lowest energy for that chosen structure. Thus the energy spectrum of all the CSA structures is obtained by threading the designed sequence against all 103 346 structures.

b. Propensity score. Propensity score is developed to test the performance as a cumulative index across all the chosen structures by matching the predicted ranks of the vertices with the probability of finding for H or P in native sequences (S_n). This parameter is evaluated in two different ways. (a) Site-wise propensity—the propensity for a lattice point (Fig. 1) to be occupied by an H or a P residue, and (b) Propensity for rank ordered sites to occupy an H or a P residue. In case (a), the score is calculated as follows: There is one designed sequence (S_d or S_a or S_e) for each of the thirteen structures. These thirteen sequences (of any one of the three types) are threaded into their own target structures on the lattice (Fig. 1). The number of times a lattice site is occupied by P-type residue is counted (n_c) and the fraction of times ($n_c/13$) it is found is called the site-wise propensity of that site. Same procedure is also followed for the native sequences (S_n), except for the possibility of more than one sequence for a given structure. The results of site-wise propensity evaluation are presented in Table IV.

In case (b) the vertices are ordered according to the evaluated ranks (R_e, R_d, and R_a in Table II). The corresponding native sequences are also ordered according to ranks. Now the fraction of times residue P takes up a particular rank ordered site in the native sequences S_n is evaluated. This procedure is followed for all the thirteen structures. Cumulative propensity for a rank-ordered site for a P residue is obtained by averaging over all the thirteen structures. The results obtained for the designed sequences S_e, S_d, and S_a are plotted in Fig. 4.

III. RESULTS AND DISCUSSION

Three different types of sequences were predicted for all thirteen structures [Table I] as described in the method section and the details of evaluation are shown for structure 4 in Table II. The sequences S_d explicitly take into account both the degree and position of vertex in the graph. The sequence S_a basically takes into account the degree information, slightly biasing its position in the graph by reverse ordering among vertices with same degree as described earlier. The sequence S_e considers the position of the vertices as obtained by only Hvc and does not explicitly consider the degree information. The performance of the designed sequences as evaluated in the sequence and the structure spaces are presented below.

A. Performance in the sequence space

The energy distribution of sequences in composition $10H$-$17P$ and $12H$-$15P$ was analyzed for all the structures by generating all possible sequences. A typical energy distri-

TABLE III. Eluding sequences (%).

Structure	S_d	S_a	S_e
	(a) Composition: 10H-17P		
1	0.000 00	0.001 33	0.076 02
2	0.000 01	0.003 35	0.074 00
3	0.000 00	0.001 64	0.001 52
4	0.001 64	0.004 31	0.001 64
5	0.000 00	0.001 73	0.003 14
6	0.000 01	0.001 73	0.003 14
7	0.000 00	0.005 19	0.000 71
8	0.000 05	0.010 22	0.000 05
9	0.000 05	0.001 96	0.143 08
10	0.000 05	0.001 94	0.033 50
11	0.000 00	0.002 19	0.010 76
12	0.000 00	0.000 97	0.010 91
13	0.000 24	0.000 96	0.001 74
	(b) Composition: 12H-15P		
1	0.000 01	0.005 53	0.138 46
2	0.000 20	0.005 44	0.135 26
3	0.000 00	0.011 37	0.005 44
4	0.000 20	0.021 17	0.000 20
5	0.000 00	0.005 47	0.007 20
6	0.000 24	0.005 48	0.007 21
7	0.000 00	0.023 03	0.002 75
8	0.000 62	0.009 78	0.000 62
9	0.000 05	0.002 62	0.004 64
10	0.000 69	0.002 62	0.004 56
11	0.000 01	0.004 61	0.018 36
12	0.000 00	0.004 74	0.032 39
13	0.000 03	0.004 45	0.002 51

TABLE IV. Propensity of vertices to prefer $H(=0)$ or $P(=1)$ in the native and predicted sequences for 12H-15P composition.

	Lattice site			
Number	S_d	S_a	S_e	S_n
1	1.00	0.69	1.00	0.92
2	0.85	0.54	0.69	0.60
3	1.00	0.92	1.00	0.98
4	0.38	0.77	0.31	0.68
5	**0**	**0**	0.15	**0**
6	0.38	0.62	0.38	0.47
7	**1.00**	**1.00**	**1.00**	**1.00**
8	0.38	0.85	0.23	0.50
9	**1.00**	**1.00**	**1.00**	**1.00**
10	0.85	0.46	0.77	0.62
11	**0**	**0**	0.46	**0**
12	0.77	0.46	0.69	0.65
13	**0**	**0**	0.31	**0**
14	**0**	**0**	**0**	**0**
15	**0**	**0**	0.15	**0**
16	0.85	0.38	0.62	0.70
17	**0**	**0**	**0**	**0**
18	0.38	0.85	0.31	0.55
19	**1.00**	0.92	**1.00**	**1.00**
20	0.54	0.69	0.38	0.61
21	1.00	0.85	1.00	0.97
22	0.77	0.54	0.77	0.53
23	**0**	**0**	0.08	**0**
24	0.54	0.54	0.46	0.75
25	**1.00**	**1.00**	**1.00**	**1.00**
26	0.31	0.92	0.23	0.48
27	**1.00**	**1.00**	**1.00**	**1.00**
Correlation[a]	*0.95*	*0.90*	*0.87*	⋯

[a]Correlation w.r.t. native sequence(s); bold numbers show complete agreement with propensity of native sequences(s).

bution [for structure 4] curve is given in Fig. 3. The energies of the native sequence(s) (S_n) and the designed sequences S_d, S_a, and S_e are also indicated in the figure. Generally, the energies of the designed sequences are close to that of S_n indicating a reasonably good performance of all the designed sequences at gross level. The percentage of eluding sequences for all the chosen structures from different designed sequences for the two compositions are given in Table III. The percentage of eluding sequences is very small for all the predicted sequences, indicating that all of them do reasonably well in sequence space. Furthermore, the eluding sequences from the S_d sequences are extremely small and are zero in many cases, i.e., sequence predicted is one of the native sequences (S_n). This indicates that our procedure, which takes into account the information on the vertex, its degree and its position in global context, has performed very well.

The performance of the designed sequences is also evaluated by the propensity measure described in the method section. Propensities are measured in two different ways. Table IV lists the propensities of the lattice points (Fig. 1) for the preference of an H [=0] or a P [=1] type of residue for the native and the designed sequences. This analysis illustrates that there are some vertices on the lattice which have implicit preference for an H or a P residue irrespective of the conformation and the propensities of other vertices are conformation dependent. The native sequences S_n clearly show the features of the lattice structure in which a propensity value of zero is obtained for the vertices of high degree—the body centered vertex [14] and six face centered vertices [5, 11, 13, 15, 17, and 23]. The corner vertices [7, 9, 19, 25, and 27] of degree one show propensity values very close to one. A slight deviation from the propensity value of unity in the other corner vertices 1, 3, and 21 is due to the fact that they have a degree of two because of the occupancy of terminal residues in certain structures. Twelve edge-centered vertices of degree two are degenerate from the point of view of lattice structure. For a composition of 12H-15P, five of these sites have to be occupied by H type of residues and the selection of these five sites is conformation specific. Propensity wise these vertices take up values ranging from 0.5 to 0.75. Now a comparison shows that the lattice features [a low-propensity value for vertices of high degree and vice versa] exhibited by the natural sequences S_n are also exhibited by the designed sequences, which accounts for a good performance of all the designed sequences at the gross level. The sequences S_e which do not explicitly take into account the degree information, however, do not reproduce the complete preference for H type residue for high degree vertices. The correlation coefficient between the site propensities of S_n and designed sequences show that the designed sequences S_d perform better than S_a or S_e. A better performance of S_d over S_a in discriminating the twelve vertices of degree two clearly indicates that the input of the vertex position from graph theoretical method has been valuable in designing a good sequence.

The propensity values are also presented (see Fig. 4) according to the rank ordered vertices of designed sequences for a composition of $12H$-$15P$. Here we see that the seven vertices of high degree are clearly selected as the top ranking centers by the designed sequences S_a and S_d, preferring H type of residue. Similarly, six of the terminal vertices with degree one emerge as the low ranking centers with a preference for P residue. The designed sequences S_e, although exhibit similar propensities, a small deviation is observed because of a lack of explicit input of degree information in deciding the vertex weight. Further, the edge-centered vertices of degree two are well discriminated by the 3D conformation on the lattice by the sequences S_d. This can be seen by low propensities of the rank ordered vertices 8, 9, and an increasing propensity from 0.5 to 1.0 for rank ordered vertices 14–21. The performance of the sequences S_a is particularly bad for rank ordered vertices 8–20. This is expected as most of these vertices are of degree two, which need discrimination and we have deliberately reversed the rank ordering. It is interesting to note that the performance of S_e for centers 7–20 is in-between that of S_a and S_d confirming the importance of Hvc in proper rank ordering. A good performance of the designed sequences by the present method of rank ordering vertices can perhaps become more dramatic when bigger lattices with more degenerate vertices of high degree are used and sequences containing more than two types of amino acid residues are considered, since unique ranks are obtained for each vertex of a chosen structure by this procedure.

B. Performance in structure space

The probability of a good designed sequence in the sequence space becoming a good sequence also in the structure space is high. However, there is no guarantee that other structures will not have lower energy than the chosen target structure. Hence the performance of the designed sequences are also evaluated in structure space of all CSA structures. The results are presented in Fig. 5 and Table V for compositions (a) $10H$-$17P$ and (b) $12H$-$15P$. The symbols E_{d-0}, E_{a-0}, E_{e-0}, and E_{n-0} in Fig. 5 and Table V correspond to the energy of the sequences S_d, S_a, S_e, and S_n respectively in their target structures. The subscript ''min (max)'' stands for the lowest (highest) energy that was obtained by threading the concerned sequence into all (103,346) CSA structures. The energies of sequences S_d, S_a, S_e, and S_n appear in the lower part of the energy spectrum in Fig. 5 [$E_{\max}$ is shown on the upper part of the figure for comparison] assuring that all the designed sequences do reasonably well in identifying their target structure as one of the low-energy structures. The mean energy for the nontarget structures in Table I was 22.1 (standard deviation +/− 3.5) for $10H$-$17P$ [27.0 (standard deviation ±3.6) for $12H$-$15P$] indicating the diversity of the chosen structures in terms of their energies. In concurrence with the previous section, the sequences S_d have performed extremely well with the energies of some of the S_d sequences being as good as those of S_n in both the chosen compositions as seen in Fig. 5. The result presented in Table V gives the performance of all the predicted sequences in a quantitative way. The value of (E_{d-0} $-E_{d-\min}$) being zero in all the cases shows that the sequences S_d indeed predict the target structure as native state in all the structures and in both the compositions. Sequences S_a and S_e on the other hand recognize the target structures as the native states only in about 50% of the structures and have found structures which are better than the target by about 0.3–0.9 units of energy in other cases. The energy difference for the target structure from the predicted and the native sequences [$(E_{d-0}-E_{n-0})$, etc.] given in Table V also shows that the S_d sequences are close to the native sequences than the other two designed sequences.

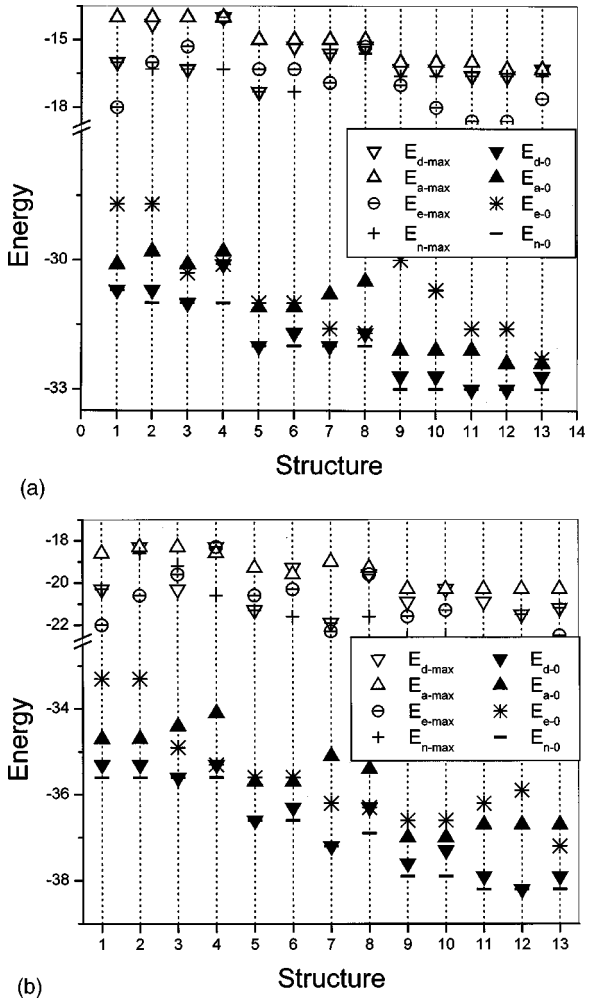

FIG. 5. The energy spectrum in the conformational space of all CSA conformations obtained by the native sequences (S_n) and the designed sequences S_d, S_a, and S_e generated for thirteen structures given in Table I for the compositions-(a) $10H$-$17P$ and (b) $12H$-$15P$. $E_{d-0}(E_{a-0},E_{e-0},E_{n-0})$ is the energy of the designed sequence $S_d(S_a,S_e,S_n)$ in target conformation. $E_{d-\max}(E_{a-\max},E_{e-\max},E_{n-\max})$ is the maximum energy of designed sequence $S_d(S_a,S_e,S_n)$ in the entire conformation space.

Thus from the present investigation of HP model on a $3\times3\times3$ lattice it is possible to design good sequences which are energetically stable and whose minimum energy conformation is the target structure, provided the composition of the amino acids and their interaction potentials are given. It has been shown that the connectivity of the nodes and their position in the global context of the protein structure play an

TABLE V. Performance of the predicted sequences in the structure and sequence spaces in terms of energy differences.[a]

Structure	$E_{d-0}-E_{d-\min}$	$E_{a-0}-E_{a-\min}$	$E_{e-0}-E_{e-\min}$	$E_{d-0}-E_{n-0}$	$E_{a-0}-E_{n-0}$	$E_{e-0}-E_{n-0}$
			Composition: $10H$-$17P$			
1	0	0.6	0.6	0	0.6	2.0
2	0	0.6	0.6	0.3	1.2	2.3
3	0	0	0	0	0.9	0.7
4	0	0	0	0.9	1.2	0.9
5	0	0.3	0.3	0	0.9	1.0
6	0	0	0	0.3	0.9	1.0
7	0	0	0	0	1.2	0.4
8	0	0	0	0.3	1.5	0.3
9	0	0	0	0.3	0.9	3.0
10	0	0.3	0.3	0.3	0.9	2.3
11	0	0	0	0	0.9	1.4
12	0	0	0	0	0.6	1.4
13	0	0	0	0.3	0.6	0.7
			(b) Composition: $12H$-$15P$			
1	0	0.6	0.3	0.3	0.9	2.3
2	0	0.6	0	0.3	0.9	2.3
3	0	0.6	0	0	1.2	0.7
4	0	0.9	0	0.3	1.5	0.3
5	0	0.3	0	0	0.9	1.0
6	0	0.3	0.3	0.3	0.9	1.0
7	0	0	0	0	2.1	1.0
8	0	0.9	0	0.6	1.5	0.6
9	0	0	0.3	0.3	0.9	1.3
10	0	0	0.3	0.6	0.9	1.3
11	0	0	0	0.3	1.5	2.0
12	0	0	0.3	0	1.5	2.3
13	0	0	0	0.3	1.5	1.0

[a]$E_{d-0}(E_{a-0},E_{e-0},E_{n-0})$ is the energy of the designed sequence $S_d(S_a,S_e,S_n)$ in target conformation. $E_{d-\min}(E_{a-\min},E_{e-\min},E_{n-\min})$ is the minimum energy of designed sequence $S_d(S_a,S_e,S_n)$ in the entire conformation space.

important role in deciding the nature of the sequence of amino acids for a desired protein structure.

IV. CONCLUSIONS

The noncovalent interactions in polymer structure on a $3\times3\times3$ lattice were represented in a Laplacian matrix form. The roots of these set of simultaneous equations (eigenvalues) and their components (eigenvectors) are used to derive topological indices for a chosen conformation (graph). A unique weight is assigned for all the residues (vertices) of the polymer, which are used to design the sequence that takes up the desired conformation as its native state.

The designed sequences are evaluated in the sequence and structure spaces. The performance evaluation shows that lattice structure itself decides the occupancy of an H or a P type residue on certain vertices. However, there are several vertices which are conformational dependent in their choice of H or a P residue. It is demonstrated that the weights derived by the degree of the vertex, its position in the global topology as evaluated by the highest eigenvalue and its vector components, perform extremely well in designing sequences for a chosen structure. The designed sequences reproduce the residue preferences indictated by the general lattice structure as well as those dependent on the conformation. The designed sequences also perform well in the structure space in predicting the chosen conformation as the native state.

We realize that our method, which derives information from the native state topology, is only one of the ways of approaching the problem of sequence design. The possibility that other schemes similar in spirit performing equally well or better than our scheme, however, is not ruled out. The success of our scheme encourages generalization of its application to bigger lattices with more than two types of amino acid residues and perhaps to real proteins.

ACKNOWLEDGMENTS

We thank Super Computer Education and Research Center (SERC) and BioInformatics Center of the Indian Institute of Science for computational facilities. One of us (B.S.S.) thanks Council for Scientific and Industrial Research (CSIR) for research fellowship.

[1] M. H. Cordes, A. R. Davidson, and R. T. Sauer, Curr. Opin. Struct. Biol. **6**, 3 (1996).
[2] B. I. Dahiyat and S. L. Mayo, Science **278**, 82 (1997).
[3] T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, and D. C. Richardson, Proc. Natl. Acad. Sci. U.S.A. **91**, 8747 (1994).
[4] A. Lombardi, J. W. Bryson, and W. F. DeGrado, Biopolymers **40**, 495 (1996).

[5] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[6] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[7] W. F. DeGrado, Z. R. Wasserman, and J. D. Lear, Science **243**, 622 (1989).
[8] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).
[9] J. W. Ponder and F. M. Richards, J. Mol. Biol. **193**, 775 (1987).
[10] C. Pabo, Nature (London) **301**, 200 (1983).
[11] R. Kaul and P. Balaram, Bioorg. Med. Chem. Lett. **7**, 105 (1997).
[12] K. E. Drexler, Proc. Natl. Acad. Sci. U.S.A. **78**, 5275 (1981).
[13] S. F. Betz, J. W. Bryson and W. F. DeGrado, Curr. Opin. Struct. Biol. **5**, 457 (1995).
[14] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **80**, 2237 (1998).
[15] D. T. Jones, Curr. Opin. Biotechnol. **6**, 452 (1995).
[16] G. Tuchscherer and M. Mutter, J. Pept. Sci. **1**, 3 (1995).
[17] J. R. Desjarlais and N. D. Clarke, Curr. Opin. Struct. Biol. **8**, 471 (1998).
[18] D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).
[19] M. H. Hao and H. A. Scheraga, Curr. Opin. Struct. Biol. **9**, 184 (1999).
[20] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[21] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).
[22] C. Micheletti, F. Seno, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **80**, 2237 (1998).
[23] S. M. Patra and S. Vishveshwara, Biophys. Chem. **84**, 13 (2000).
[24] N. Kannan and S. Vishveshwara, J. Mol. Biol. **292**, 441 (1999).
[25] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000).
[26] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **82**, 3372 (1999).
[27] D. Baker, Nature (London) **405**, 39 (2000).
[28] N. Trinajstic, *Chemical Graph Theory* (CRC, Boca Raton, FL, 1992).
[29] I. Gutman and O. E. Polansky, *Mathematical Concepts in Organic Chemistry* (Springer, Berlin, 1986).
[30] S. C. Basak, A. T. Balaban, G. D. Grunwald, and B. D. Gute, J. Chem. Info. Comp. Sci **40**, 891 (2000).
[31] D. R. Flower, Protein Eng. **7**, 1305 (1994).
[32] I. Koch, F. Kaden, and J. Selbig, Proteins: Struct., Funct., Genet. **12**, 314 (1992).
[33] P. J. Artymiuk, D. W. Rice, E. M. Mitchell, and P. Willett, Protein Eng. **4**, 39 (1990).
[34] T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[35] I. Bahar, A. R. Atilgan, R. L. Jernigan, and B. Erman, Proteins: Struct., Funct., Genet. **29**, 172 (1997).
[36] M. C. Demirel, A. Atilgan, R. L. Jernigan, B. Erman, and I. Bahar, Protein Sci. **7**, 2522 (1998).
[37] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[38] S. M. Patra and S. Vishveshwara, Int. J. Quantum Chem. **71**, 349 (1999).
[39] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[40] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[41] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98**, 7420 (1993).
[42] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[43] O. Ivanciue and A. T. Balaban, *Encyclopedia of Computational Chemistry, Vol. 2*, edited by P. V. R. Schleyer (Wiley, New York, Singapore, 1998), p. 1174.
[44] I. Gutman and D. Cvetkovic, Croat. Chem. Acta **49**, 115 (1977).