Topological Shape and Size of Peptides: Identification of Potential Allele Specific Helper T Cell Antigenic Sites

Chandan Raychaudhury,*,§ Asok Banerjee,† Partha Bag,‡ and Syamal Roy⊥

Computer Division and Department of Immunology, Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Jadavpur, Calcutta 700032, India, Department of Biophysics, Bose Institute, Calcutta 700054, India, and CompuVision, 102 Rishi Bankim Chandra Road (Extn.), Calcutta 700028, India

Received April 23, 1998

A database of primary sequences of 28 immunogenic peptides, known to elicit T cell response, derived from five different haplotypes was compiled to identify allele specific helper T cell antigenic sites using a rule based graph-theoretical method. The prediction was based on the identification of allele specific patterns in the form of "topological shape and size" present in the peptides. Indices computed from weighted connected graph models of amino acid side chains and peptides were used in this purpose. The system was trained by 10 A^d and 10 non-A^d restricted peptide sequences, assigned actives and inactives, respectively, chosen randomly from the database, and four A^d and four non-A^d restricted sequences were kept as test peptides. This allowed the system to learn about "topological shape and size" specific for A^d restricted peptides from the differences, if any, they had with the inactive peptides in that respect. The system made 100% correct prediction for the training set peptides and misclassified only one inactive peptide of the test set. The system also identified crucial residues for lambda repressor 12-24 and insulin A-chains. This identification also shows that activity related/crucial residues could be located at varying distances from the peptide terminals. To our knowledge, the method is unique of its kind in the literature and may find application in the rational design of synthetic vaccines and other peptides of immunological importance.

1. INTRODUCTION

Identification and characterization of T-cell antigenic sites have received substantial attention in recent years for both its practical usefulness such as to design synthetic peptide vaccines[1−3] particularly for deadly diseases like AIDS where use of attenuated live virus vaccine or even killed virus vaccine may be extremely risky[3] and also to address problems of immunological phenomena.[4] It is well-known that helper T-cells recognize proteolytically trimmed fragments of protein antigens from diverse origin in the context of polymorphic class II MHC molecules to induce immune response.[5,6] However, structural properties playing a role in peptide-MHC interactions to elicit T cell response has to be explored systematically in order to identify potential T cell epitopes.

Few methods have been proposed to predict T helper cell antigenic sites from primary sequences of protein antigens exploiting physical/physicochemical properties of the amino acids.[7−10] Though all these methods have successes, they have limitations too. The general validity of the methods of Margalit et al.[7] and Rothbard and Taylor,[8] based on amphipathic α helix and common sequence pattern, respectively, are controversial as it has been pointed out that nearly the entire sequence of a protein may have T cell epitopes depending on the studied haplotype.[11] Thus any algorithm would pick up some T cell target sequences correctly if it is not done with respect to a given restriction.[9,11]

Rothbard and Taylor[8] found some indication of the presence of allele specific pattern in T cell epitopes, and Guillet et al.[9] and Altuvia et al.[10] later on developed algorithms to predict allele specific T cell determinants by identifying motifs in the sequences. However, since binding residues may be located at varying distances from the peptide terminals for class II associated peptides,[12] emphasis should be given not only on the structural properties of the residues but also on the identification of activity related individual residues for a better understanding of the role a residue plays in peptide-MHC interaction. This, in turn, might help rational design of useful peptides with precision.

Determination of the X-ray structures of class I human (HLA-A2)[13] and class II human (HLA-DR1)[14] molecules helps to look into the problem more specifically and carry out a systematic study. In the case of HLA-DR1,[14] peptides bind in extended conformation onto the cleft of MHC molecules with a contact length of 15 residues.[14] This clearly shows that fitting is extremely important and a necessary criterion to get T cell response from peptide-MHC interactions. Thus attention should be paid to take into account the size as well as the shape of amino acid side chains in developing an algorithm such that potential epitopes can be screened from protein sequences with high accuracy.

Again, since many residues in the cleft of the MHC molecules vary among alleles,[12] presence of allele specific shape and size in the antigenic determinants is a distinct possibility, and algorithms should be developed for searching such patterns. However, topological structures of the side

* Corresponding author. E-mail: nandy@x400.nicgw.nic.in. FAX: 91-33-473-5197.
† Bose Institute.
‡ CompuVision.
§ Computer Division, Indian Institute of Chemical Biology.
⊥ Department of Immunology, Indian Institute of Chemical Biology.

10.1021/ci980052w CCC: $18.00
Published on Web 01/05/1999

Table 1. Physical and Topological Descriptors of Leucine and Isoleucine: A Comparative View

amino acid	sum of the atomic weights of side chain atoms	M^{S}_{S} value of the side chain
leucine	57.1111	19.6944
isoleucine	57.1111	22.3611

chains and peptides should be considered to accomodate flexibility in the measures of shape and size since a large number of peptides from diverse origin dock onto a single peptide binding domain of MHC molecules. Graph-theoretical methods[15] may be conveniently used in this purpose.

In this paper, we propose, for the first time, a graph-theoretical approach for the prediction of allele specific helper T-cell antigenic sites. An index M^{S}_{S}, reflecting topological shape and size of amino acid side chains, has been developed for the present study. This index along with another graph-theoretical index D^{-4}, the distance exponent index,[16] and a rule based system[17,18] have been used to identify activity related residues which could be involved in the formation of allele specific topological shape and size in immunogenic peptides. The work has been carried out on a database of primary sequences of 28 helper T cell target sequences derived from five different haplotypes. Only those sequences have been considered which are known to activate T cells i.e., the T cell motifs. The system has identified all the peptides of the training set correctly and mispredicted only one peptide of the test set. The system has also picked up experimentally determined crucial residues of lambda repressor 12-24 and insulin A-chains satisfactorily. It appears from all these findings that the proposed method may find application in the rational design of synthetic vaccines and peptides relevant to studies with T cell epitopes as well as other immunological phenomena of importance.

2. METHODOLOGY

For the present study a new graph-theoretical index M^{S}_{S}, reflecting topological shape and size of amino acid side chains, has been used. The index is computed from the weighted connected graph model of amino acid side chains. Attention has been paid such that values for two side chains are not identical. To have this, weighting factors have been used at the atomic level instead of considering gross molecular descriptor(s). For example, as shown in Table 1, the algebraic sum of the atomic weights of the side chains of leucine and isoleucine are the same. However, the index M^{S}_{S}, which is also defined on atomic weights, has different values for the two molecules. The index has been computed for the side chains of 20 standard amino acids. These 20 (distinct) values (given in Table 2) have been used as weighting factors in another connected graph model of amino acid sequence to compute another index D^{-4}, the distance exponent index,[16] for each Cα vertex of the graphs representing the peptides under consideration. These D^{-4} values are used to identify activity related residues and predict allele specific determinants using a rule based system.[17,18]

2.1. Side Chain Index. Let G be a rooted weighted graph model for the side chain of histidine (H) where each vertex of G other than the root vertex (which is actually a Cα vertex) is weighted by the atomic weight (taking oxygen = 16.00) of the corresponding atom in histidine (Figure 1). One

Table 2. M^{S}_{S} Values of the Side Chains of 20 Standard Amino Acids

serial no.	side chain of amino acid	M^{S}_{S} value
1	alanine	13.0200
2	valine	21.0267
3	leucine	19.6944
4	isoleucine	22.3611
5	proline	19.0242
6	phenylalanine	23.5288
7	tryptophan	26.1975
8	methionine	21.9200
9	glycine	1.0100
10	serine	18.3533
11	threonine	22.3567
12	cysteine	23.7067
13	tyrosine	24.1954
14	asparagine	21.8567
15	glutamine	19.9689
16	aspartic acid	22.0200
17	glutamic acid	20.0233
18	lysine	18.9756
19	arginine	19.0434
20	histidine	23.5283

can readily see in Figure 1 that several paths are connecting the vertices of G with its root vertex (encircled), and the weighted vertices are lying on these paths. It is also clear that the paths are branched in all the vertices of G other than the pendant vertices. Now, more distant a vertex is from the root vertex, longer is the path between them. In fact there may be more than one path connecting two vertices and also there may be more than one path of the same topological distance between the two vertices. However, for our purpose, we consider those paths between two vertices which are the shortest ones, and we consider the sum of the weights of the vertices on those paths in computing the index. If there are more than one path of the same topological distance, then that path is considered the sum of the weights of whose vertices is minimum. Regarding branching of the path at the nonpendant vertices, we assign a probability value to each branched path emerging from a vertex. For example, if there are two paths proceeding from a vertex we assign 1/2 to each edge emerging from that vertex. If an edge entering a vertex has a weight, say 1/2, and three paths go out of that vertex, then each edge coming out of the vertex would be assigned a probability value 1/6 (= 1/3 × 1/2). The process would be continued until one reaches the last edge of a path.

It may be noted that the paths are considered to be directed from the root vertex to the pendant vertices although that has not been indicated in G. However, the pendant vertices in G are labeled as 1, 2, ..., 6, the other vertices as 7, 8, ..., 12, and the root vertex (encircled) as 0, and the probability values are also given on the edges (Figure 1). Clearly, the probability value would become lower if a path contains more branched vertices. It may also be noted that in a cyclic graph where pendant vertices do not exist, one may delete certain edges to get a tree graph in the process of obtaining paths between the root and other vertices (in a tree, the path between two vertices is unique) for the sake of computational advantage. Now, considering the probability values assigned to the edges connected to the pendant vertices and the vertex weights (atomic weight in our case) in the shortest paths connecting the root vertex with the pendant vertices we

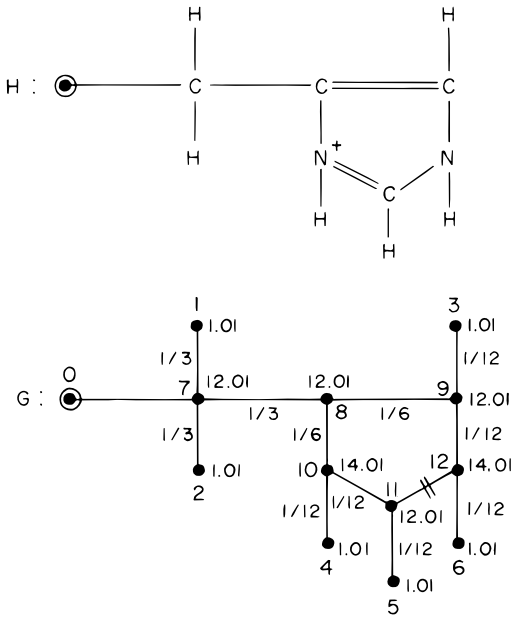

Figure 1. Side chain of histidine (H) and its weighted molecular graph (G). The edge marked || is not considered for the computation of graph-theoretical index.

define an index $M^{S}{}_{S}$ related to molecular shape and size as follows.

Let a path P in a graph G' be connecting the root vertex to a pendant vertex v and the edge connected to v has a probability value p. Then we compute the path value $P^{v}{}_{p}$ of the path P as

$$P^{v}{}_{p} = p(w_1 + w_2 + ... + w_r) \qquad (1)$$

where w_i, $i = 1, 2, ..., r$ are the weights of the vertices in P other than the root vertex which is not weighted. Thus, if there are h such paths

$$P_1, P_2, ..., P_h$$

in G' connecting h pendant vertices of G' to the root vertex and

$$P^{v1}{}_{p1}, P^{v2}{}_{p2}, ..., P^{vh}{}_{ph}$$

be the path values then we define the molecular shape and size related index $M^{S}{}_{S}$ for the root vertex as

$$M^{S}{}_{S} = \sum_{j=1}^{h} P^{vj}{}_{pj} \qquad (2)$$

To illustrate the procedure, the value of the index $M^{S}{}_{S}$ for the root vertex of G may, thus, be obtained from its path values. The path P^{l} connecting the root (0) and the pendant vertex 1 is P^1 = 0-7-1, where the numbers correspond to the vertices in the path. Similarly, P^2 = 0-7-2; P^3 = 0-7-8-9-3; P^4 = 0-7-8-10-4; P^5 = 0-7-8-10-11-5; P^6 = 0-7-8-9-12-6.

Hence, after putting the probability values and vertex weights we get the path values for G following eq 1 as

$$P^{l}{}_{1/3} = P^{2}{}_{1/3} = 1/3(12.01 + 1.01)$$

$$P^{3}{}_{1/12} = 1/12(12.01 + 12.01 + 12.01 + 1.01)$$

$$P^{4}{}_{1/12} = 1/12(12.01 + 12.01 + 14.01 + 1.01)$$

$$P^{5}{}_{1/12} = 1/12(12.01 + 12.01 + 14.01 + 12.01 + 1.01)$$

$$P^{6}{}_{1/12} = 1/12(12.01 + 12.01 + 12.01 + 14.01 + 1.01)$$

Therefore, $M^{S}{}_{S}$ value for the root vertex v_0 of G (Figure 1) may be calculated from these path values using eq 2 and, thus

$$M^{S}{}_{S}(v_0) = 23.5283$$

It may be noted that in getting the shortest paths between the root vertex and the pendant vertices 5 and 6, edge (11, 12) in G is not required and hence may be deleted or suppressed (as indicated in Figure 1) for the sake of computational ease as explained earlier.

2.2. Distance Exponent Index. The distance exponent index D^{x} [16] for the Cα vertices of the peptides are computed considering the graph model G'' of amino acid sequences as shown in Figure 2. The weights W_1, W_2, ..., W_n in G'' are $M^{S}{}_{S}$ values of the amino acid side chains and hence depend on the kind of residue present at that position. The distance exponent index D^{x} for a Cα vertex in the graph model of a peptide is computed using the following equation

$$D^{x}(C\alpha) = \sum_{s=1}^{n} W_s \times (d_s)^{x} \qquad (3)$$

where d_s is the topological distance of the sth weighted vertex from the Cα vertex under consideration, $s = 1, 2, ..., n$, and x may take any real value. For example, in G'', the D^{-4} value (taking $x = -4$) of Cα2 is

$$D^{-4}(C\alpha 2) = [W_1 \times 2^{-4}] + [W_2 \times 1^{-4}] + [W_3 \times 2^{-4}] + ... + [W_{n-2} \times (n-3)^{-4}] + [W_{n-1} \times (n-2)^{-4}] + [W_n \times (n-1)^{-4}]$$

As an illustration, D^{-4}(Cα2) in the peptide 12-24 of lambda repressor protein (Figure 3) i.e., D^{-4} value of the Cα vertex for glutamic acid (E) at the second position from the left in the peptide is

$$(19.6944 \times 2^{-4}) + (20.0233 \times 1^{-4}) + (22.0200 \times 2^{-4}) + (13.0200 \times 3^{-4}) + (19.0434 \times 4^{-4}) + (19.0434 \times 5^{-4}) + (19.6944 \times 6^{-4}) + (18.9756 \times 7^{-4}) + (13.0200 \times 8^{-4}) + (22.3611 \times 9^{-4}) + (24.1956 \times 10^{-4}) + (20.0233 \times 11^{-4}) + (18.9756 \times 12^{-4}) = 22.9304$$

This may be verified from Table 6 with the value given against residue no. 2. It may be noted that as the negative exponent is increased, the index focuses more on the close neighbors, and the effect of the distant neighbors decreases rapidly depending on the negative exponent considered. It has been found[16] that negative exponents -3 or -4 are very

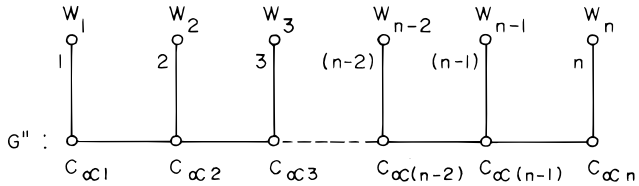

Figure 2. Graph-theoretical model of peptide chain considering C vertices and side chain weights.

Amino acid sequence of lambda repressor 12–24	C_α position in graph model	M^S_S value of side chains
L	C_{α_1}	19.6944
E	C_{α_2}	20.0233
D	C_{α_3}	22.0200
A	C_{α_4}	13.0200
R	C_{α_5}	19.0434
R	C_{α_6}	19.0434
L	C_{α_7}	19.6944
K	C_{α_8}	18.9756
A	C_{α_9}	13.0200
I	$C_{\alpha_{10}}$	22.3611
Y	$C_{\alpha_{11}}$	24.1956
E	$C_{\alpha_{12}}$	20.0233
K	$C'_{\alpha_{13}}$	18.9756

Figure 3. Amino acid sequence with C vertices and attached weights of the side chains in the graph model of the peptide 12-24 in lambda repressor protein.

optimum and serve the purpose quite satisfactorily. In the present study, we have considered -4 as the negative exponent that is D^{-4} index has been computed for the Cα vertices.

3. DATABASE

The database for the present study, compiled from the literature, is composed of known primary sequences of 28 immunogenic peptides (Table 3). Fourteen of them (1−14) are restricted by A^d molecule, six of them (15−20) are restricted by A^b molecules, three each (21−23 and 24−26) by A^k and E^k molecules, and two (27 and 28) by H-2^k molecules, respectively.

4. ORDERING AND RULES

Once a database is formed, few peptides which are antigenic in some particular strain are called active peptides, and those which are antigenic in other strains are called inactive peptides. The D^{-4} index for all the Cα vertices of all active and inactive peptides of a training set, created by random selection of sequences from the database, are then computed by a computer program. When all of these D^{-4} values are arranged in a nondecreasing order (i.e., in the ordering, a value would be the same or greater than the immediately preceeding value), several ranges of D^{-4} values are detected in the ordering.

Values coming from active peptides are said "active ranges" and those contributed by inactive peptides are said "inactive ranges". However, certain rules are followed for the selection of active and inactive ranges:[18]

1. Three or more consecutive values coming exclusively from active peptides or exclusively from inactive peptides are considered to form an active range or an inactive range, respectively.

2. Some single value coming from both active and inactive peptides is not considered to be from a range by itself or together with other values unless more than two-thirds of the peptides contributing this value are of same activity. However, that single value, if it does not form a separate range, may be included in an active or an inactive range which may form independently following (1).

There may have been several regions in the ordering where no ranges are found.

Once the active and inactive ranges are obtained, the activity of a peptide is evaluated on the basis of the occurrence of D^{-4} values, computed for the Cα vertices of its graph model, in active and inactive ranges. A peptide is predicted to be active if the index values of all or some of its atoms fall

1. ONLY IN THE ACTIVE RANGE(S) OR
2. IN BOTH ACTIVE AND INACTIVE RANGES AND THE NUMBER OF INDICES IN THE ACTIVE RANGES IS MORE THAN THAT IN THE INACTIVE RANGES.

Otherwise, the peptide is predicted to be inactive.

5. RESULTS

In the database (Table 3) we assigned 14 peptides for A^d bearing mouse as active and the remaining 14 peptides derived from four other haplotypes as inactive. Subsequently, sequences of 10 active and 10 inactive peptides were chosen randomly to form a training set, and those of remaining four active and four inactive peptides were kept back as a test set. When the system was run with the training set sequences, it learned about allele specific topological shape and size present in active and inactive peptides in terms of M^S_S values (Table 2) capable of discriminating closely related structures (Table 1) and D^{-4} values.

When all the D^{-4} values computed for the training set peptides were put into order (materials and methods), active and inactive ranges were found in the ordering. Prediction of activity of each peptide, belonging to the training set and the test set, was then carried out on the basis of the occurrences of their D^{-4} values, in active and inactive ranges. It was found that activities of all the 20 peptides of the training set were predicted correctly, and only one A^k restricted inactive peptide, hen lysozyme 46-60, of the test set was misclassified. The evaluation of activity of all the peptides is given in Table 4. Some of the ranges identified in the ordering of D^{-4} values have been shown in Table 5, and activity prediction has been illustrated in Table 6 taking lambda repressor 12-24 (no. 7, Table 3) as an example.

It was also found that the system could identify crucial residues, determined experimentally, quite efficiently. The finding for lambda repressor 12-24 is shown in Table 7. It is also clear from Table 7 that activity related/crucial residues,

Table 3. Twenty-Eight Known Antigenic Peptide Sequences and Their MHC Restrictions

serial no.	peptide name	amino acid sequence	ref
	(a) A^d Restricted Peptides		
1	bovine insulin B-chain 5-15	HLCGSHLVEAL	8
2	cytochrome bovine 13-25	KCAQCHTVEKGGK	8
3	chicken ovalbumin 323-339	ISQAVHAAHAEINEAGR	8
4	bovine insulin A-chain 1-21	GIVEQCCASVCSLYQLENYCN	22
5	sperm whale myoglobin 106-118	FISEAIIHVLHSR	8
6	staph nuclease 61-80	FTKKMVENAKKIEVEFDKGQ	8
7	lambda repressor 12-24	LEDARRLKAIYEK	8
8	flu haemaglutinin 130-142	HNTNGVTAACSHE	8
9	sheep insulin A-chain 1-21	GIVEQCCAGVCSLYQLENYCN	22
10	porcine insulin A-chain 1-21	GIVEQCCTSICSLYQLENYCN	22
11	equine insulin A-chain 1-21	GIVEQCCTGICSLYQLENYCN	22
12	HSV_sd 245-260	APYSTLLPPELSETP	25
13	[Ala 14] insulin A-chain 1-14	GIVEQCCASVCSLA	22
14	[Ala 12/13] insulin A-chain 1-14	GIVEQCCASVCAAY	22
	(b) A^b Restricted Peptides		
15	cytochrome horse 45-58	GFTYTDANKNKGIT	8
16	hen egg lysozyme 78-93	IPCSALLSSDITASVN	8
17	staph nuclease 91-110	YIRADFKMVNEALVRQGLAK	8
18	pigeon cytochrome C 45-58	GFSYTDANKNKGIT	8
19	lamda repressor 73-88	VEEFSPSIAREIYEMY	8
20	herp. glycoprotein D 1-20	KYALADASLKMADPNRFRGK	8
	(c) A^k Restricted Peptides		
21	hen lysozyme 46-60	NTDGSTDYGILQINS	8
22	hen lysozyme 34-45	FESNFNTEATNR	8
23	malaria circums. protein 326-343	PSDKHIEQYLKKIKNSIS	8
	(d) E^k Restricted Peptides		
24	sp. whale myoglobin 69-78	LTALGAILKKK	8
25	staph nucl. 81-100	RTDKYGRGLAYIYADGKMVN	8
26	cytochrome moth 89-103	NERADLIAYLKQATK	8
	(e) H-2^k Restricted Peptides		
27	hepat. B suface antigen 140-154	TKPSDGNCTCIPIPS	8
28	hepat. B pre S 120-132	MQWNSTTFHQTLQ	8

picked up by the system, are not consecutive. It has also been observed that the system can identify crucial residues (1, 2, 3, 4, 5) for insulin A-chains some of which were in the training set and the others in the test set.

6. DISCUSSION

In the present study, a rule based graph-theoretical system has been used for the identification of allele specific helper T cell antigenic sites in terms of the identification of allele specific topological shape and size present in the peptides. The results (Table 4) indicate that the predictive power of the system is significantly high producing 100% correct prediction for the training set containing 20 peptides and only one misprediction for the test set of eight peptides. This seems to indicate that topological shape and size play a vital role in the fitting[19] of peptides onto MHC cleft for T cell activation.

Moreover, the approach is an open ended one in that one does not train the system to match with some known pattern. The system rather identifies the activity related topological shape and size from the presence of this structural characteristics in the peptides of the training set.

Therefore, it is likely that the system is capable of identifying topological shape and size and the associated residues in peptides responsible for interacting with polymorphic MHC molecules for the prediction of allele specific epitopes. The only incorrect prediction is for an inactive peptide, the hen lysozyme 46-60 (peptide no. 21 of the test set, Table 4).

It is also apparent from the data given in Table 7 that activity related residues may not necessarily be consecutive to form a motif. This is in agreement with the findings that, for class II associated peptides crucial residues may be located at varying distances from the ends of the peptides.[12] It is also interesting to note that the present system has been able to identify all three residues 19, 21, and 24 of lambda repressor 12-24 (Table 7) reported to be responsible for the immunogenicity of this peptide.[21] The system has also picked up two other residues, 16 and 23, as related to activity besides the above mentioned three (Table 6). It is possible that residues 16 and 23 help strengthen the fitting of the peptide onto the cleft of MHC molecules. This is further substantiated by the findings for insulin A-chains. Experiments suggest that the first five residues of insulin A-chain 1-21 and insulin A-chain 1-14 and its truncated analogs are crucial,[22] and our system has identified those five residues as related to activity in all the insulin A-chains studied be in the training set or the test set.

Results also indicate that both $M^S{}_S$ and D^{-4} have played useful roles for the system to learn about similarities and differences in topological shape and size among A^d and non-A^d restricted peptides in the process of identifying allele specific patterns. In computing $M^S{}_S$, atomic weights have been used to take care of size effect of the constituent atoms in the side chains. Though molecular weight is commonly used as a bulk steric parameter mimicking molecular size,[23] consideration of atomic weight along with skeletal branching, on which three-dimensional structure of a molecule depends,[24] is expected to relate $M^S{}_S$ to the shape and size of

Table 4. Evaluation of the Activity of Antigenic Peptides Considering A^d Restricted Peptides as Active and A^b, A^k, E^k, and H-2^k Restricted Peptides as Inactives

serial no.	peptide as in Table 3	activity[a] assigned	activity[a] predicted
	Training Set		
1	1	+	+
2	2	+	+
3	3	+	+
4	4	+	+
5	5	+	+
6	6	+	+
7	7	+	+
8	8	+	+
9	9	+	+
10	10	+	+
11	15	−	−
12	16	−	−
13	17	−	−
14	18	−	−
15	20	−	−
16	22	−	−
17	23	−	−
18	24	−	−
19	25	−	−
20	27	−	−
	Test Set		
1	11	+	+
2	12	+	+
3	13	+	+
4	14	+	+
5	19	−	−
6	21	−	+[b]
7	26	−	−
8	28	−	−

[a] + = active (antigenic in A^d bearing mouse). − = inactive (antigenic in strains other than A^d bearing mouse). [b] This is the only misclassified peptide.

Table 5. Formation of Ranges in the Ordering of D^{-4} Values of Cα Vertices in the Peptides of the Training Set (Table 4)

serial no. in the ordering	D^{-4} (Cα) value	peptide no. (residue no. in the peptide)	activity[a] and range type
95	21.7980	7 (8)	+ active range
96	21.8021	7 (5)	+
97	21.8151	6 (10)	+
98	21.8292	8 (11)	+
99	21.8345	9 (12)	+
100	21.8641	8 (13)	+
101	21.9022	4 (12)	+
104	21.9433	17 (16)	− inactive range
105	21.9672	27 (14)	−
106	21.9844	24 (9)	−
107	21.9858	27 (3)	−
108	21.9871	17 (3)	−
133	22.6098	3 (3)	+ not a range
134	22.7006	3 (14)	+
135	22.7161	20 (9)	−
136	22.7256	7 (7)	+
137	22.7477	16 (7)	−

[a] + = value contributed by active peptide. − = value contributed by inactive peptide.

flexible or topological structure of a molecule. The reason is that atomic weight is a very fundamental physical property to be used to take care of the relative effects of molecular size as precisely as possible which is very essential for the present purpose. At the same time, since the sum of the probability values used to characterize skeletal branching is always 1 for any structure, it appears that the differences in shape of the molecules would be taken care of by the index on the basis of the differences in the probability values assigned to each path of a molecular graph.

Table 6. Activity Prediction for the Peptide Lambda Repressor 12-24 (LEDARRLKAIYEK)

serial no.	D^{-4} (Cα) value	residue no.	range type	prediction
1	16.2832	4	not in a range	
2	16.3859	9	not in a range	
3	20.6688	13	in active range	
4	21.3341	1	not in a range	
5	21.7980	8	in active range	
6	21.8022	5	in active range	active
7	22.1110	6	not in a range	
8	22.7256	7	not in a range	
9	22.9304	2	not in a range	
10	23.1130	12	in active range	
11	24.7019	3	not in a range	
12	25.3806	10	in active range	
13	27.3776	11	not in a range	

Table 7. Amino Acids Found as Critical Residues in the Peptide 12-24 (LEDARRLKAIYEK) of Lambda Repressor Protein from Experimental Studies (Ref 21) and by the Present Molecular Topology Method

residue number in the peptide	
obtained from experimental studies	obtained by the present method
19, 21, 24	16, 19, 21, 23, 24

On the other hand, D^{-4} includes the effects of the close side chains more than that of the distant side chains for a given Cα vertex. Hence, it is believed that active ranges reflect regions of similar topological shape and size in peptides restricted by a specified haplotype (A^d in our case). Thus such ranges, in turn, help identify activity related residues since D^{-4} is computed for Cα vertex representing the position of a residue. However, that D^{-4} includes the effects of all the side chains in a peptide at a given residue position seems to be an important aspect taken care of by the index since peptides bind with class II MHC molecules in extended conformation with full contact length.[14]

Therefore, since the evaluation of activity by the present system is made on the basis of important structural characteristics like topological shape and size of amino acid side chains as well as on the basis of the position of individual amino acid in a peptide, it is hoped that the present intelligent system would help identify, and perhaps more certainly for initial screening of, allele specific helper T cell target sequences with precision and would guide design synthetic vaccines and other peptides of interest through identification and replacement of residues.

ACKNOWLEDGMENT

We are thankful to the late Prof. B. K. Bachhawat, Department of Biochemistry, University of Delhi, New Delhi and Prof. A. N. Bhaduri, Indian Institute of Chemical Biology, Calcutta, for their constant encouragement in carrying out this work. Thanks are also due to the late Prof. P. K. Bhattacharyya for his helpful comments during the course of this work. We are extremely thankful to Prof. A. J. Hopfinger and the referees for their suggestions which

helped us to make improvements in the manuscript. Financial assistance received from the Council of Scientific & Industrial Research and the Indian Council of Medical Research, New Delhi, India, is also thankfully acknowledged.

REFERENCES AND NOTES

(1) Berzofsky, J. A.; Cease, K. B.; Spouge, J. L.; Margalit, H.; Berkower, I. J.; Good, M. F.; Miller, L. H.; DeLisi, C. *Immunol. Rev.* **1987**, *98*, 9.
(2) Good, M. F.; Berzofsky, J. A.; Miller, L. H. *Ann. Rev. Immunol.* **1988**, *6*, 663.
(3) Berzofsky, J. A. *J. Clin. Invest.* **1988**, *82*, 1811.
(4) Werdelin, O.; Mouritsen, S.; Patersen, B. L.; Sette, A.; Buus, S. *Immunol. Rev.* **1988**, *106*, 181.
(5) Rothbard, J. B.; Gefter, M. L. *Ann. Rev. Immunol.* **1991**, *9*, 527.
(6) Schartz, R. H. *Ann Rev. Immunol.* **1985**, *3*, 237.
(7) Margalit, H.; Spouge, J. L.; Cornett, J. L.; Cease, K. B.; DeLisi, C.; Barzofsky, J. A. *J. Immunol.* **1987**, *138*, 2213.
(8) Rothbard, J. B.; Taylor, W. R. *EMBO J.* **1988**, *7*, 93.
(9) Guillet, J.-G.; Hoebeke, J.; Lengagne, R.; Tate, K.; Borras-Herrera, F.; Strosberg, A. D.; Borras-Cuesta, F. *J. Mol. Recognition* **1991**, *4*, 17.
(10) Altuvia, Y.; Berzofsky, J. A.; Rosenfeld, R.; Margalit, H. *Mol. Immunol.* **1994**, *31*, 1.
(11) Roy, S.; Scherer, M. T.; Briner, T. J.; Smith, J. A.; Gefter, M. L. *Science* **1989**, *244*, 572.
(12) Engelhard, V. H. *Ann. Rev. Immunol.* **1994**, *12*, 181.
(13) Bjorkman, P. J.; Saper, M. A.; Samraoui, B.; Bennett, W. S.; Strominger, J. L.; Wiley, D. C. *Nature* **1987**, *329*, 512.
(14) Brown, J. H.; Jardetzky, T. S.; Gorga, J. C.; Stern, L. J.; Stern, R. G.; Urban, R. G.; Strominger, J. L.; Wiley, D. C. *Nature* **1993**, *364*, 33.
(15) Harary, F. *Graph Theory*; Addison-Wesley, Reading, MA, 1972.
(16) Raychaudhury, C.; Klopman, G. *Bull. Soc. Chim. Belg.* **1990**, *99*, 255.
(17) (a) Klopman, G.; Raychaudhury, C. *J. Comput. Chem.* **1988**, *9*, 232. (b) Klopman, G.; Raychaudhury, C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 12.
(18) Raychaudhury, C.; Dey, I.; Bag, P.; Biswas, G.; Das, B. N.; Roy, P. K.; Banerjee, A. *Arzneim.-Forsch./Drug Res.* **1993**, *43*(2), 1122.
(19) Marx, J. *Science* **1995**, *267*, 459.
(20) Sette, A. S.; Buss, S.; Appella, E.; Smith, J. A.; Chesnut, R.; Miles, C.; Colon, S. M.; Grey, H. M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 3296.
(21) Guillet, J.-G.; Lai, M.-Z.; Briner, T. J.; Smith, J. A.; Gefter, M. L. *Nature* **1986**, *324*, 260.
(22) Willims, D. B.; Ferguson, J.; Gariepy, J.; McKay, D.; Teng, Y-T.; Iwasaki, S.; Hozumi, N. *J. Immunol.* **1993**, *151*, 3627.
(23) Dearden, J. C. In *Practical Applications of Quantitative Structure−Activity Relationship (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990.
(24) Silipo, C.; Vittoria, A. In *Comprehensive Medicinal Chemistry, Vol. 4: Quantitative Drug Design*; Hansch, C., Sammes, P. G., Taylor, J. B., Ramsden, C. A., Eds.; Pergamon Press: Oxford, U.K., 1990.
(25) Sette, A. S.; Buus, S.; Colon, S.; Miles, C.; Grey, H. M. *J. Immunol.* **1988**, *141*, 45.

CI980052W