The EMBO Journal vol.7 no.1 pp.93 - 100, 1988

A sequence pattern common to T cell epitopes

Jonathan B.Rothbard and William R.Taylor[1,2]

Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX and [1]Birkbeck College, London WC1E 7HX, UK
[2]Present address: National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Communicated by M.J.Crumpton

An analysis of the known cytotoxic and helper T cell epitopes has revealed similarity within their primary sequences. These similar motifs, characteristic of the known determinants, have been incorporated into predictive templates that have been used successfully to define eight helper and three cytotoxic epitopes in four different proteins. When the defined epitopes are segregated by restriction element, allele specific subpatterns emerge centering around the general pattern. The presence of similarities argues that the binding of peptide antigens to class I and class II is similar in nature. In addition, these motifs can be used to predict accurately areas within proteins capable of being recognized by individual MHC class I and class II molecules.

Key words: T cell epitopes/MHC protein/peptides/T cell

Introduction

As the molecular basis of recognition of protein antigens by the immune system has been elucidated, the different mechanisms used by B and T lymphocytes have become more apparent. Historically, the two populations of lymphocytes could be distinguished by their ability to bind protein antigens. T cells are unable to bind antigens in the absence of histocompatibility proteins, whereas B cells specifically bind antigens via their surface immunoglobulins in a manner characteristic of other ligand–receptor interactions on eukaryotic cell surfaces (reviewed by Schwartz, 1985). The differences can now be understood due to the recent demonstration that instead of binding proteins with their native structure intact, both helper and cytotoxic T cells are stimulated by peptide fragments of the immunogen (Townsend *et al.*, 1986; Schwartz, 1985). A structural model for the recognition of peptides has evolved from the binding experiments which proved that both human and murine class II histocompatibility proteins specifically bind peptide antigens. Initially demonstrated using detergent solubilized class II molecules in equilibrium dialysis experiments (Babbitt *et al.*, 1985), the peptide–class II complexes have been shown to be sufficiently stable to be separated by gel filtration (Buus *et al.*, 1986; T.Jardetzky, personal communication). In the large majority of cases, the antigenic peptide was bound preferentially by the class II molecule through which T cell recognition was restricted (Buus *et al.*, 1987).

Competition between peptides restricted by the identical histocompatibility allele provided another important aspect to the current model. Initially demonstrated using cellular proliferation assays (Guillet *et al.*, 1986), and recently extended to both cytotoxicity and direct binding assays (J. Maryanski and B.Askonas, personal communication; Buus *et al.*, 1986), the inhibition argues that both class I and class II MHC molecules have a single antigen combining site. Taken together these results imply that T cell recognition involves the antigen receptor binding a complex between MHC and peptide.

Concomitant with these experiments, empirical analyses of the determinants on proteins recognized by T cells revealed that they were both quite limited in number and shared several common features. Not only were the majority of these regions distinct from areas recognized by antibodies, but a high percentage also correlated with areas that were helical in the intact protein and/or could be modelled into amphipathic helices (reviewed by Berzofsky, 1986). Rather than consider secondary structure, we have analysed the primary structure of both helper and cytotoxic determinants and have observed that a large percentage contain a linear pattern composed of a charged residue or a glycine followed by two hydrophobic residues (Rothbard, 1986). In this report we expand on this initial observation defining the general T cell pattern and examine its correlation with known epitopes, discuss its statistical significance, and review our success in using the algorithm to predict previously unidentified helper and cytotoxic determinants. Finally, by segregating the known epitopes by restriction elements and centering the epitopes on the described patterns we can identify allele specific subpatterns which are being used to examine how peptide antigens bind to both class I and class II molecules.

Results and discussion

The known helper and cytotoxic T cell epitopes which make up the database of this analysis are shown in Table I. They are a combination of human, mouse and guinea pig epitopes which have either been published or communicated to us. Each peptide has been shown to stimulate either the corresponding antigen specific helper clone, cell line, or act as a target for the correct cytotoxic T cell clone. The list has a degree of subjectivity. We have tried to minimize any prejudice by including only those areas of proteins which have been directly shown to be epitopes and not inferred by a process of elimination. In those cases where a variety of peptides were analysed for a response, we have tried to choose areas that stimulated at sufficiently low concentrations ($\sim 0.1-1$ μmol) that we were confident that all the required residues were included in the peptide. Finally, we have tried to be comprehensive; any deletion of a defined epitope has not been intentional.

The epitopes have been arranged to reveal a pattern within their primary structure. The pattern which we have found was determined by examining the residues which make up

Table I. Compilation of reported T cell epitopes

A. Known epitopes containing the patterns

Position	1	2	3	4	5	6	7	8	9	MHC restriction	Reference
Bovine insulin B-chain 5-15	H	L	C	G	S	H	L	V	E	A	L	A^d	Thomas *et al.* (1981)
Cytochrome moth 89-103		N	E	R	A	D	L	I	A	Y	L K Q A T K	E^b, E^k	Hedrick *et al.* (1982)
Cytochrome bovine 13-25	K	C	A	Q	C	H	T	V	E	K	G G K	A^d	Corradin *et al.* (1983)
Cytochrome horse 45-58						G	F	T	Y	T	D A N K N K G I T	A^b, A^k	Suzuki and Schwartz (1986)
Ovalbumin 323-339	I	S	Q	A	V	H	A	A	H	A	E I N E A G R	A^d	Shimonkevitz *et al.* (1984)
Flu nucleoprotein (1934/68) 335-349	S	A	A	F	E	D	L	R	V	L	S F I R G	B37	Townsend *et al.* (1986)
Flu nucleoprotein (1968) 366-379		A	S	N	E	N	M	D	A	M	E S S T L	D^b	Townsend *et al.* (1986)
Flu nucleoprotein (1934) 366-379		A	S	N	E	N	M	E	T	M	E S S T L	D^b	Townsend *et al.* (1986)
Flu haemagglutinin A/Aichi/2/68 48-66	K	I	C	N	N	P	H	R	I	L	D G I D C T L I D	Ia^k	Mills *et al.* (1986)
Flu haemagglutinin PR/8 111-120			F	E	R	F	F	E	I	F	P K E	A^d	Hackett *et al.* (1983)
Flu haemagglutinin A/Texas/1/77 114-131				A	S	S	G	T	L	E	F I N E G F N W T G V	Human	Lamb and Green (1983)
		A	S	S	G	T	L	E	F	I	N E G F N W T G V		
Flu haemagglutinin A/Texas/1/77 307-319						P	K	Y	V	K	Q N T L K L A T	DR1	Rothbard *et al.* (in press)
	P	K	Y	V	K	Q	N	T	L	K	L A T		
Sperm whale myoglobin 69-78				L	T	A	L	G	A	I	L K K	E^k	Livingstone *et al.* (1987)
Sperm whale myoglobin 106-118					F	I	S	E	A	I	I H V L H S R	A^k, A^d	Berkower *et al.* (1985)
	F	I	S	E	A	I	I	H	V	L	H S R		
Sperm whale myoglobin 110-121					A	I	I	H	V	L	H S R H P G	E^d	Livingstone *et al.* (1987)
Sperm whale myoglobin 132-146			N	K	A	L	E	L	F	R	K D I A A K Y	E^d	Berkower *et al.* (1986)
						N	K	A	L	E	L F R K D I A A K Y		
Rat myelin basic protein 5-17	P	S	Q	R	H	G	S	K	Y	L	A T A	$A^k_\alpha A^u_\beta$	Zamvil *et al.* (1986)
Ragweed allergen Ra3 51-65			E	V	W	R	E	E	A	Y	H A A D I K D	H-2 d,k,s,q	Kuisaki *et al.* (1986)
	E	V	W	R	E	E	A	Y	H	A	A D I K D		
								E	V	W	R E E A Y H A A D I K D		
Human acetyl choline receptor γ 125-142				K	S	Y	C	E	I	I	V T H F P F D Q Q N C	Rat	Lennon *et al.* (1985)
Hen egg lysozyme 34-45		F	E	S	N	F	N	T	E	A	T N R	A^k	Allen *et al.* (1984)
Hen egg lysozyme 46-61	N	T	D	G	S	T	D	Y	G	I	L Q I N S R	A^k	Babbitt *et al.* (1986)
Hen egg lysozyme 78-93	I	P	C	S	A	L	L	S	S	D	I T A S V N	A^b	Manca *et al.* (1984)
Hen egg lysozyme 81-96			S	A	L	L	S	S	D	I	T A S V N C A K	E^k, A^b	Shastri *et al.* (1985)
Herpes glycoprotein D 1-23	K	Y	A	L	A	D	A	S	L	K	M A D P N R F R G K D L P	A^b	Heber-Katz *et al.* (1986)
Staph. nuclease 61-80						F	T	K	K	M	V E N A K K I E V E F D K G Q	A^d	Finnegan *et al.* (1986)
Staph. nuclease 81-100			R	T	D	K	Y	G	R	G	L A Y I Y A D G K M V N	E^k	Finnegan *et al.* (1986)
	R	T	D	K	Y	G	R	G	L	A	Y I Y A D G K M V N		
Staph. nuclease 91-110					Y	I	Y	A	D	G	K M V N E A L V R Q G L A K	A^b	Finnegan *et al.* (1986)
	Y	I	Y	A	D	G	K	M	V	N	E A L V R Q G L A K		
			Y	I	Y	A	D	G	K	M	V N E A L V R Q G L A K		
VP1 foot-mouth virus 145-160			R	G	D	L	Q	V	L	A	Q K V A R T L P	Guinea pig	Francis *et al.* (1985)
Angiotensin II							D	R	V	Y	I H P F	Guinea pig	Thomas *et al.* (1981)
Angiotensin II (V-5)							D	R	V	Y	V H P F	Guinea pig	Thomas *et al.* (1981)
Fibrinopeptide B 5-14				D	N	E	E	G	F	F	S A R	Guinea pig	Thomas *et al.* (1980)
Minimum stimulatory polymer	tnp		E	Y	K	(tnp)		E	Y	A	E Y A	H-2d	Singh *et al.* (1980)
HLA CW3 171-182			Y	L	K	N	G	K	E	T	L Q R A	K^d	Maryanski *et al.* (1986)
Lambda repressor 12-24	L	E	D	A	R	R	L	K	A	I	Y E K	A^d	Guillet *et al.* (1986)
HIV GP 120 428-443	K	Q	I	I	N	M	W	Q	E	V	G K A M Y A	H-2 k,s,d	Cease *et al.* (1987)
HIV GP 120 112-124						H	E	D	I	I	S L W N Q S L K	H-2 s	Cease *et al.* (1987)
Epstein–Barr virus LMP 43-53				V	M	S	D	W	T	G	G A L L	A1	Thorley-Lawson and Israelson (1987)

B. Known epitopes lacking the described patterns

Rat meylin basic protein 1-11	Ac A S Q K R P S Q R H G	A^u	Zamvil *et al.* (1986)
Hepatitis B pre-S 120-132	M Q W N S T T F H Q T L Q	H-2 k,q,s	Milich *et al.* (1986)
G. Pig myelin basic protein 68-88	G S L P Q K S Q R S Q D E N P V V H F	Rat	Kibler *et al.* (1977)
Pigeon cytochrome C 45-58	G F S Y T D A N K N K G I T	A^bA^k	Suzuki and Schwartz (1986)
Rabies glycoprotein 32-44	D E G C T N L S G F S Y M	H-2 d	MacFarlan *et al.* (1984)
Hepatitis B surface antigen 95-109	L V L L D Y Q G M L P V C P L	H-2 q	Milich *et al.* (1985)
Hepatitis B surface antigen 140-154	T K P S D G N C T C I P I P S	H-2 k	Milich *et al.* (1985)
Malaria circumspor. protein 326-343	P S D K H I E Q Y L K K I K N S I S	A^k	Good *et al.* (1987)
Flu haemagglutinin 130-142	H N T N G V T A A C S H E	A^d	Hackett *et al.* (1985)

C. Epitopes successfully identified using the patterns for prediction

Flu nucleoprotein (1934/68) 50-63	S D Y E G [R L I Q] N S L T I	K^k	Bastin *et al.* (1987)
Flu nucleoprotein (1934/68) 147-158	T Y Q R T R A L V [R] T G	K^d	Taylor *et al.* (1987)
Flu matrix protein 17-31	S G P L K A E I A Q [R] L E D V	DR1	Rothbard *et al.* (in press)
Flu matrix protein 57-68	K G I L G [F] V F T L T V	A2	Gotch *et al.* (1987)
	K G I L G F V F [T] L T V		
Tuberculosis 65 kd protein 112-126	Y E K I G A E L V K [E] V A K K	Human, H-2 b,d,k	Lamb *et al.* (1987)
	Y E K I G A E L V K [E] V A K [K]		
Tuberculosis 65 kd protein 163-184	K R G I E K A V D [K] V T E T L L K D A K E V	Human	Lamb *et al.* (1987)
	K R G I E K A V D [K] V T E [T] L L K D A K E V		
	K R G I E K A V D K V T E T L L [K] D A K E V		
Tuberculosis 65 kd protein 227-243	T F G L Q L E L T E [G] M R F D K G	Human	Lamb *et al.* (1987)
Tuberculosis 65 kd protein 242-266	S G Y F V T D A E R Q E A V L [E] E P Y I L	Human	Lamb *et al.* (1987)
	K G Y I S G Y F V [T] D A E R Q E A V L E E P Y I L		
	K G Y I S [G] Y F V T D A E R Q E A V L E E		
Tuberculosis 65 kd protein 437-459	R K H R I E D A V R [N] A K A A V E E G I V A G	Human	Lamb *et al.* (1987)
	R K H R I E D A V R N A K A A V [E] E G I V A G		
Rat myelin basic protein 35-47	T G I L D S I G R F F S [G]	$E^u_\alpha E^\lambda_\beta$	Zamvil *et al.* (submitted)
Rat myelin basic protein 89-101	V [H F F K] N I V T P R T P	A^λ	Sakai *et al.* (submitted)

Amino acid compositions of different positions within defined T-cell epitopes containing proposed patterns.

position

1 8A 2V 6L I 3F 3Y M W 2T 8E 2K H 3R 4N 2Q 2S C 3G
26 hydrophobic 14 charged 10 polar 3 Gly

2 5A 5V 8I Y M 3T 4D 2E 7R K H 4N 2Q 4S 5G P
23 hydrophobic 15 charged 11 polar 5 Gly 1 Pro

3 4A 2V 9L 2I F 2Y 2M 3T 3D 4E R 4K H 3N 4Q 8S 2C 5G P
25 hydrophobic 13 charged 17 polar 5 Gly 1 Pro

4 7D 18E 14K 7H 7R 11G
53 charged 11 Gly

5 16A 9V 9L 9I 6F 5Y 4M 5T W
64 hydrophobic

6 11A 11V 13L 9I 6F 3Y 3M W 7T
64 hydrophobic

7 4A 6V 3L 2I F 4Y T 4D 10E 5K 4H 4R 4N 4Q 5S P 2G
20 hydrophobic 27 charged 14 polar 1 Pro 2 Gly

8 5A 2L 3I 3F 2Y 7T 7E 7K 3H 5R 4N Q 6S P 6G
22 hydrophobic 22 charged 11 polar 1 Pro 6 Gly

9 8A 6V 6L 4I F 2Y M W 2T 4D 5E 3K H 3R 3N Q 3S 2G 2P
30 hydrophobic 16 charged 8 polar 2 Gly 2 Pro

the epitopes as members of characteristic sets based on their physical properties (Taylor, 1986). In this classification some amino acids are members of more than one group, for example both threonine and tyrosine are classified as both hydrophilic and hydrophobic due to the physical properties of their side chains. The legend to Table I lists the amino acid compositions of the nine positions within the known epitopes. Of the 57 epitopes listed, 48 have within them a region of sequence composed of (i) a charged residue or glycine, followed by (ii) two hydrophobic residues. In the 48 epitopes there are 64 examples of this three amino acid pattern. There are 47 examples where the residue in the next position was either charged or polar. In the 17 cases where the fourth position was a hydrophobic residue, each had a polar residue in the next position. Those peptides which have more than a single pattern have been listed accordingly (Table IA). At this time there are nine known T cell epitopes which lack the described patterns and are separately shown in Table IB. The 11 epitopes which have been predicted successfully using these motifs are listed in Table IC.

Even though the number of defined T cell epitopes still is quite small, several generalizations can be made: (i) unlike the majority of B cell epitopes, the vast majority of T cell epitopes are composed of residues corresponding to a linear sequence in the intact protein, (ii) the minimum length necessary for either full stimulation of helper T cells or maximum lysis in cytotoxic T cell recognition is 8–12 amino acids, (iii) the particular epitope recognized is dependent on the MHC protein, yet there are areas of globular proteins which a variety of clones of different histocompatibility types recognize, but often with different fine specificities (reviewed by Berzofsky, 1986; Rothbard, 1986; Livingston and Fathman, 1987).

To determine whether these patterns had any importance we have (i) attempted to calculate the statistical significance of these occurrences, (ii) examined the literature to determine whether any of these epitopes lacking residues composing the pattern or with mutations which have the pattern could be recognized by T cells, and (iii) used the template to identify previously unidentified helper and cytotoxic T cell epitopes.

The statistical significance of the correlation of the patterns with known epitopes was analysed by calculating the probability of the pattern of 4 (charged or glycine, hydrophobic, hydrophobic, polar or glycine), and 5 amino acids (charged or glycine, hydrophobic, hydrophobic, hydrophobic or proline, polar or glycine) to occur at least once in a random 15 amino acid peptide (Table II). As can be seen, in each case the pattern has occurred about twice as often as expected. However, we must be cautious in the interpretation of the statistics both because the pattern was initially determined from an examination of the epitopes themselves and because protein sequences are not random. To partially address the latter fact, we have also calculated the probability of occurrence in a 15 amino acid peptide whose composition reflects the average composition of known proteins (Dayhoff, 1976). In this case the occurrence of the pattern also occurs more often than expected.

The importance of the proposed patterns within T cell epitopes was emphasized when the experimental results which initially defined the size and fine specificity of some of the T cell determinants were reviewed (summarized in Table III). Because the minimum size of a T cell epitope is 8–12 amino acids in size, no pattern of four or five residues will be sufficient for recognition. In addition, because there are seven defined T cell epitopes which lack either pattern, such sequences cannot be absolutely required. However, an examination of the literature has revealed interesting and supportive correlations. In all cases where the minimal sequence recognized by a T cell clone has been examined, the pattern has always been intact (Table III). For example, when histidine 327 was deleted from the ovalbumin epitope the peptide lost all stimulatory activity (Watts *et al.*, 1986). In addition, when histidine 330 was modified to alanine, which removes the proposed pattern, the peptide was not recognized by the T cell hybridoma. Recent experiments examining a large number of point mutations within this epitope have independently confirmed that the proposed pattern of His-Ala-Ala-His is critical for recognition (Sette *et al.*, 1987). In the myoglobin epitope, 132–146, the deletion of glutamic acid 137 removes both the pattern and the peptide's stimulatory activity (Berkower *et al.*, 1986). Lamb and his colleagues have recently shown that lysine at 308 in the influenza haemagglutinin is necessary for recognition of their human helper clone (Rothbard *et al.*, in press). This residue also constitutes the first position of the proposed pattern. The cytotoxic epitope corresponding to residues 365–379 of influenza nucleoprotein also is interesting. Townsend and colleagues (Bastin *et al.*, 1987) elicited strain specific clones to the identical region of the protein which differ in two positions, but both contain a pattern of a charged residue,

Table II. Statistical significance of the occurrence of the linear patterns

Sets of amino acids composing the proposed patterns

1.	D E H K R G	6/20 = 0.30 charged + glycine
2.	A V L I F M W T Y	9/20 = 0.45 hydrophobic
3.	A V L I F M W T Y P	10/20 = 0.50 hydrophobic + proline
4.	D E H R K N Q S T G	10/20 = 0.50 polar + glycine

P_4 = probability of the pattern of four to occur in a random tetramer
= (charged + Gly) (phobic) (phobic) (polar + glycine)
= 0.030 (= 0.033)*

P_5 = probability of the pattern of five to occur in a random pentamer
= (charged + Gly) (phobic) (phobic) (phobic + Pro) (polar + Gly)
= 0.015 (=0.016)*

Probability of the patterns to occur in a random 15 amino acid peptide is calculated as follows:

Probability of pattern = P_n
Number of possible locations of the pattern in the peptide = $N-n+1$
where, N = number of residues in the peptide = 15
n = number of residues in the pattern = 4 or 5

Probability of pattern not to occur at a position = $1-P_n$
Probability of pattern not to occur at all possible positions = $(1-P_n)^{N-n+1}$
Probability of the pattern to occur = $1-(1-P_n)^{N-n+1}$

	If random	Actual occurrence
Tetramer	30.6% (33%)*	46/57 = 81%
Pentamer	15.3% (16%)*	18/57 = 32%

By a χ^2 test all the observed frequencies occur with a probability, $P < 0.001$

*Values in parentheses are calculated using the average amino acid composition of known protein residues (Dayhoff, 1976)

Table III. Compilation of the experimental results examining the minimum length of peptides recognized by T cell clones and hybridomas

1. Sperm whale myoglobin (Livingstone *et al.*, 1987; Berkower *et al.*, 1985, 1986)

69 a 78		
L T A L G A I L K K	+++[b]	
– – – – – G – – – – [c]	–	
106 118		
E F I S E A I I H V L H S R	+++	
– – – – – D – – – – – – – – – –	–	
110 121		
A I I H V L H S R H P G	+++	
I I H V L H S R H P G	++	
I H V L H S R H P G	+	
H V L H S R H P G	–	
A I I H V L H S	–	
A I I H V L H S R	+	
A I I H V L H S R H P	++	
132 146		
N K A L E L F R K D I A A K Y	+	
K A L E L F R K D I A A K Y	+	
A L E L F R K D I A A K Y	+	
L E L F R K D I A A K Y	+	
E L F R K D I A A K Y	+	
L F R K D I A A K Y	–	

2. Influenza nucleoprotein (Townsend *et al.*, 1986)

335 349		
S A A F E D L R V L S F I R G	+++	
A A F E D L R V L S F I R G	++	
A F E D L R V L S F I R G	+	
F E D L R V L S F I R G	+	
E D L R V L S F I R G	–	
366 379		
A S N E N M D A M E S S T L	+++	
S N E N M D A M E S S T L	++	
N E N M D A M E S S T L	++	
E N M D A M E S S T L	+	
– – – – – – E T – – – – – –	–	
– – – – – – E A – – – – – –	–	
– – – – – – D T – – – – – –	+	

3. Ovalbumin (Watts *et al.*, 1986)

323 339		
I S Q A V H A A H A E I N E A G R	+++	
A V H A A H A E I N E A G R	++	
A A H A E I N E A G R	–	
A V H A A A E I N E A G R	–	

4. Lysozyme (Allen *et al.*, 1984)

46 61		
N T D G S T D Y G I L Q I N S R	+	
G S T D Y G I L Q I N S R	+	
D Y G I L Q I N S R	+	
Y G I L Q I N S R	–	
D R Q – – – – – – F – – – – –	–	

5. Influenza haemagglutinin (Rothbard *et al.*, in press)

306 320		
C P K Y V K Q N T L K L A T G	+++	
P K Y V K Q N T L K L A T G	+++	
K Y V K Q N T L K L A T G	++	
Y V K Q N T L K L A T G	–	

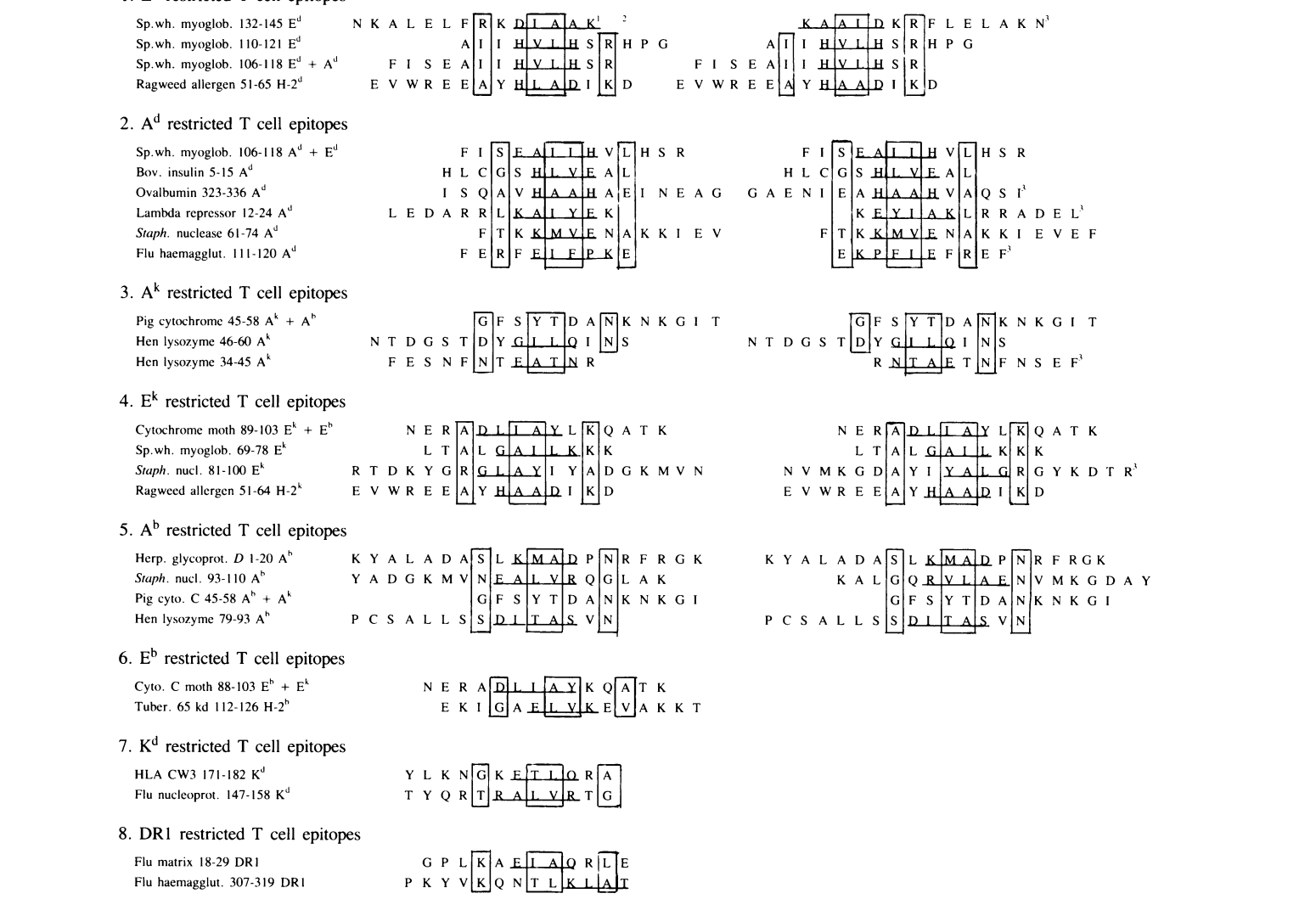
[a]Proposed patterns enclosed in boxes.
[b]Relative ability of peptide to be recognized by appropriate T cell. For details see original reference.
[c]Indicates identical residues in both sequences.

two adjacent hydrophobic amino acids followed by a polar residue. The hybrid peptides have been synthesized and tested and reveal that the glutamic acid–aspartic acid change is critical for the specific recognition and not the alanine–threonine difference (see Table III).

Atassi has examined the ability of T cells isolated from four strains of mice immunized with ragweed allergen Ra3 to respond to peptides spanning its sequence (Lennon *et al.*, 1985). In each case, the principal response was directed to a peptide corresponding to residues 51–65. This peptide contains three separate patterns. We propose that the multiple patterns allow the peptide to be recognized by four separate histocompatibility types. We would predict that each allele could recognize this peptide with different fine specificity.

As interesting as the previous analyses may be, the primary purpose of examining the known epitopes was to determine whether a predictive motif was present. The most stringent test for the merit of the proposed model is its ability to predict previously undefined T cell epitopes. From the amino acid compositions of the known epitopes we have constructed a template which identifies the described patterns in the primary sequence of a protein and orders them based on their similarity to previously identified epitopes. If a pattern of 4 amino acids is identified, we synthesize a 16 amino acid peptide which has the pattern flanked by six residues on either side. This guarantees the peptide will contain all

Table IV. Known T cell epitopes segregated by restriction element and arranged to suggest possible characteristic subpatterns



[1]Underlined residues compose the 4 and 5 amino acid patterns.
[2]Residues enclosed in boxes correspond to the two central hydrophobic amino acids of the described pattern and flanking positions which would be adjacent if the peptide adopted a helical conformation.
[3]Sequence reversed.

possible decamers containing the pattern. Similarly, with a pattern of 5 amino acids, a 15 amino acid peptide is prepared. We have successfully used the motifs to identify both helper and cytotoxic epitopes in five separate proteins, including the influenza nucleoprotein (Bastin *et al.*, 1987; Taylor *et al.*, 1987), rat myelin basic protein (Zamvil *et al.*., 1986, Zamvil *et al.*, submitted; Sakai *et al.*, submitted), influenza matrix protein (Gotch *et al.*, 1987; Rothbard *et al.*, in press) and the 65 kd protein of tuberculosis (Lamb *et al.*, 1987).

The existence of patterns within the primary sequence of T cell epitopes listed in Table I is surprising particularly because the epitopes have not been segregated by species or by class I or class II restriction. Given our understanding of how antibodies bind antigens, a representative list of B cell determinants would be unlikely to contain a consensus pattern in their primary structure because of the diversity of the antibody combining site. If we assume that the antigen receptor on the T cell is equally diverse, then any structural pattern must arise from the involvement of the MHC protein. Whether this is due solely to interactions between the epitope and the MHC molecule or whether some residues of the pattern also bind the antigen receptor cannot be determined unequivocally at this time. A third possibility is that the residues composing the pattern allow the epitope to adopt a conformation which enables the peptide to be bound both by the MHC and the T cell receptor. This latter possibility is consistent with proposals published by Berzofsky and his colleagues (DeLisi and Berzofsky, 1986, Spouge *et al.*, 1987).

In the development of these ideas we have not considered peptide conformation primarily because both helper and cytotoxic T cells recognize linear determinants and if there are similar features among these sequences, then they will be apparent in the primary structure of proteins. In addition, the ability to predict secondary structure from primary sequence is quite difficult. Even though both the pattern of 4 and 5 amino acids can be present in an amphipathic helix, they also exist in β-strands within proteins whose three-

dimensional structure has been determined.

The recent demonstrations that peptide antigens specifically bind MHC class II proteins strongly argues that at least some residues of the epitope must associate with the binding site of the restriction element (Babbitt *et al.*, 1985; Buus *et al.*, 1986). Consequently, there should be allele specific subpatterns apparent when the epitopes are segregated by restriction element. If true, then there should be similarities in the flanking residues as previously noted in some murine class II restricted epitopes (Guillet *et al.*, 1987). When the known epitopes are segregated by restriction element and aligned by the two hydrophobic residues of the described pattern, there are interesting similarities within the groups (Table IV). Including the two hydrophobic amino acids, positions with either a single or limited number of residues occur at relative locations 1, 4, 5 and 8. These residues would be juxtaposed if the peptides adopt a helical conformation, forming a conserved face of the helix, as demonstrated for the two DR1 restricted determinants (Figure 1). The similarities can be improved if selected sequences are reversed, as would effectively occur if the epitope bound to the MHC protein in the opposite direction. Such similarities can be seen in both class I and class II restricted determinants, suggesting that peptide antigens will bind to each clss of MHC molecules in a similar fashion. We should stress because of the possible occurrence of hydrophilic residues at positions 1 and 8 that such a helix need not be amphipathic. Also, not all determinants can be aligned in this manner.

This model is consistent with our recent experimental results examining DR1 restricted T cell recognition of protein antigens (Rothbard *et al.*, in press). In these experiments we defined two DR1 restricted determinants in influenza, one in the matrix protein, the second in the haemagglutinin (Table I). To investigate which residues interacted with DR1 and which were bound by the antigen receptors of the two clones, we synthesized a series of peptides composed of residues of one determinant substituted for the amino acids of the other based on the alignment shown in Table IV and Figure 1. For both the matrix and haemagglutinin specific T cell clones we were able to generate a hybrid peptide that was stimulatory. In each case six residues, corresponding to the upper façades of the two helices in Figure 1, were necessary for recognition. The alignments shown in Table IV also reveal a similarity between the two DR1 determinants and the defined E^k restricted sequences. The similarity was shown to be functionally significant by two separate experiments, the presentation of DR1 epitopes to human T cells by fibroblasts expressing I-E^k and the recognition of H-2^k restricted determinants by DR1 human lymphocytes. L cells expressing a hybrid class II molecule, composed of the α chain of E^k and the β chain of DR1 were shown to be capable of presenting each peptide to the appropriate T cell clone. In addition, L cells expressing the E^k dimer were able to present the matrix peptide to the matrix specific human helper T cell clone. We also demonstrated that peripheral blood lymphocytes from DR1 positive individuals that recognize the intact Ra3 protein of ragweed also proliferate to the ragweed peptide (Table I) previously shown to be stimulatory in H-2^k mice (Kuisaki *et al.*, 1986). Finally we have used the residues enclosed in boxes in Figure 1, which compose a face of a helix, for prediction to identify a previously undefined DR1 determinant in the 19 kd protein

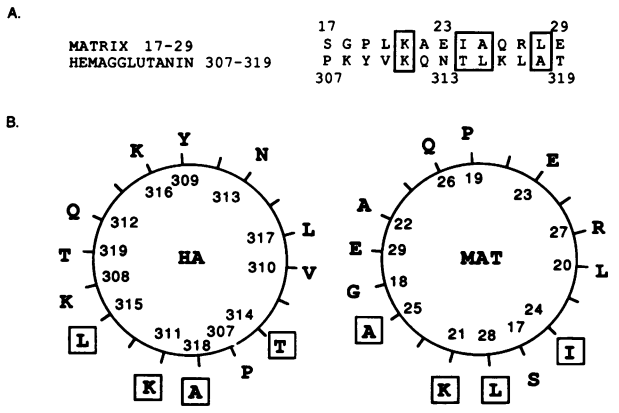


Fig. 1. Alignment of the two known DR1 restricted T cell epitopes based on two adjacent hydrophobic amino acids. Residues enclosed in boxes (**A**) are juxtaposed when the peptide adopts a helical conformation; shown when the sequence is oriented in a helical wheel (**B**).

of tuberculosis (Lamb *et al.*, submitted).

Independently, Grey and his colleagues have concluded from their experiments examining the binding of an ovalbumin peptide to I-A^d that the pattern of 4 amino acids, His, Ala, Ala, His, is critical for recognition (Sette *et al.*, 1987). However, they believe the peptide binds the MHC molecule in an extended conformation. Even though our proposed allele specific subpatterns predict that the majority of determinants will bind as a helix, we must stress that this is not an absolute requirement. Because immunological recognition is characterized by a diversity of responses and involves intermediate binding affinities, relative to the endocrine system, significant latitude in the requirements for binding are not unexpected. Consequently, we feel that both the general pattern of 4 and 5 amino acids and the allele specific subpatterns, which are apparent in the primary structure of the antigen, will be more useful than empirical predictions based solely on secondary structure.

Our empirical approach to examining the known epitopes for structural similarities has been most valuable because it has provided a basis for experimental design. Our model has evolved as more information has been obtained. We were initially impressed with the presence of the three amino acid pattern in the known determinants. For statistical reasons we included the fourth residue, which is clearly more variable, and we purposely avoided considering secondary structure. The use of the patterns of 4 and 5 amino acids allowed us to expand the known list of determinants; in particular, they helped us to identify the A2 and DR1 determinants that are currently being used in crystallographic experiments. As more epitopes have been defined, allele specific subpatterns are becoming apparent which are guiding us in analyses of how the peptides are bound by class I and class II molecules. In addition, if valid, they could explain why the linear motifs have been so useful. The critical feature of the linear and nonlinear motifs might be the adjacent hydrophobic amino acids which are present in both. Their binding to a hydrophobic area of both class I and class II molecules could be the initial binding event.

Although the extent of the diversity of peptide–MHC interactions remains to be determined, the success of these motifs in both prediction of previously undefined determinants and the identification of critical residues involved in

binding to the MHC proteins argues that for each MHC allele a specific subpattern will be identified. Consequently, regions capable of binding each histocompatibility type will be apparent from the primary structure of a protein which will greatly aid in the logical development of subunit vaccines.

Acknowledgements

We would like to thank Dr Alain Townsend, Dr Andrew McMichael, Dr Ita Askonas, Dr Jonathan Lamb and Mark Edwards for their collaboration and critical comments on the ideas written in this paper.

References

Allen,P., Strydon,D. and Unanue,E. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2489–2493.

Babbitt,B., Allen,P., Matsueda,G., Haber,E. and Unanue,E. (1985) *Nature*, **317**, 359–361.

Babbitt,B., Matsueda,G., Haber,E., Unanue,E. and Allen,P. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 4509–4513.

Bastin,J., Rothbard,J., Davey,J., Jones,I. and Townsend,A. (1987) *J. Exp. Med.*, **155**, 1508–1523.

Berkower,I., Kawamura,H., Matis,L. and Berzofsky,J. (1985) *J. Immunol.*, **135**, 2628–2634.

Berkower,I., Buckenmeyer,G. and Berzofsky,J. (1986) *J. Immunol.*, **136**, 2498–2503.

Berzofsky,J. (1986) *The Year in Immunology*. Karger, Basel, Vol. 2, pp. 28–38.

Buus,S., Sette,A., Colon,S.M., Jenis,D.M. and Grey,H. (1986) *Cell*, **47**, 1071–1077.

Buus,S., Sette,A., Colon,S., Miles,C. and Grey,H. (1987) *Science*, **235**, 1353–1358.

Cease,K., Margalit,H., Cornette,J., Putney,S., Robey,W., Ouyang,C., Streicher,H., Fischinger,P., Gallo,R., Delisi,C. and Berzofsky,J. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4249–4253.

Corradin,G., Juillerat,M., Vita,C. and Engers,H. (1983) *Mol. Immunol.*, **20**, 763–768.

Dayhoff,M.O. (1976) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC. Vol. 5, Suppl. 2, p. 301.

DeLisi,C. and Berzofsky,J. (1986) *Proc. Natl. Acad. Sci. USA*, **82**, 7048–7052.

Finnegan,A., Smith,M., Smith,J., Berzofsky,J., Sachs,D. and Hodes,R. (1986) *J. Exp. Med.*, **164**, 897–910.

Francis,M.J., Fry,C., Rowlands,D., Brown,F., Bittle,J., Houghton,R. and Lerner,R. (1985) *J. Gen. Virol.*, **66**, 2347–2354.

Good,M., Maloy,W., Lunde,M., Margalit,H., Cornette,J., Smith,G., Moss, B., Miller,L. and Berzofsky,J. (1987) *Science*, **235**, 1059–1062.

Gotch,F., Rothbard,J., Howland,K., Townsend,A. and McMichael,A. (1987) *Nature*, **326**, 881–882.

Guillet,J., Lai,M., Briner,T., Smith,J. and Gefter,M. (1986) *Nature*, **324**, 260–263.

Guillet,J., Lai,M., Briner,T., Buus,S., Sette,A., Grey,H., Smith,J. and Gefter,M. (1987) *Science*, **235**, 865–870.

Hackett,C., Dietzschold,B., Gerhard,W., Ghrist,B., Knorr,R., Gillessen,H. and Melchers,F. (1983) *J. Exp. Med.*, **158**, 294–302.

Hackett,C., Hurwitz,L., Dietzschold,B. and Gerhard,W. (1985) *J. Immunol.*, **135**, 1391–1394.

Hedrick,S., Matis,L., Hecht,T., Samelson,L., Longo,D., Heber-Katz,E. and Schwartz,R. (1982) *Cell*, **30**, 141–152.

Heber-Katz,E., Hollosi,M., Dietzschold,B., Hudecz,F. and Fasman,G. (1986) *J. Immunol.*, **135**, 1385–1390.

Kibler,R., Fritz,R., Chou,F., Chou,C., Peacock,N., Brown,N. and McFarlin,D. (1977) *J. Exp. Med.*, **146**, 1323–1330.

Kuisaki,J., Atassi,H. and Atassi,M. (1986) *Eur. J. Immunol.*, **16**, 236–240.

Lamb,J. and Green,N. (1983) *Immunology*, **50**, 659–666.

Lamb,J., Ivanyi,J., Rees,A., Rothbard,J., Howland,K., Young,R. and Young,D. (1987) *EMBO J.*, **6**, 1245–1249.

Lennon,V., McCormick,D., Lambert,E., Griesmann,G. and Atassi,M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 8805–8809.

Livingstone,A. and Fathman,C.G. (1987) *Annu. Rev. Immunol.*, **5**, 477–502.

MacFarlan,R., Dietzschold,B., Witkor,T., Kiel,M., Houghton,R., Lerner, R., Sutcliffe,J. and Koprowski,H. (1984) *J. Immunol.*, **133**, 2748–2752.

Manca,F., Clarke,J., Miller,S., Sercarz,E. and Shastri,N. (1984) *J. Immunol.*, **133**, 2075–2078.

Maryanski,J., Pala,P., Corradin,G., Jordan,B. and Cerottini,J. (1986) *Nature*, **324**, 578–579.

Milich,D., Peterson,D., Leroux-Roels,G., Lerner,R. and Chisari,F. (1985) *J. Immunol.*, **134**, 4203–4208.

Milich,D., McLachlan,A., Chisari,F. and Thornton,G. (1986) *J. Exp. Med.*, **164**, 532–547.

Mills,K., Skehel,J. and Thomas,B. (1986) *J. Exp. Med.*, **163**, 1477–1490.

Rothbard,J. (1986) *Ann. Inst. Pasteur*, **137E**, 518–526.

Rothbard,J., Lechler,R., Howland,K., Bal,V., Long,E., Rafek,Z., Taylor, W. and Lamb,J. (1987) *Cell*, in press.

Schwartz,R.H. (1985) *Annu. Rev. Immunol.*, **3**, 237–262.

Sette,A., Buus,S., Colon,S., Smith,J., Miles,C. and Grey,H. (1987) *Nature*, **328**, 395–399.

Shastri,N., Ohi,A., Miller,A. and Sercarz,E. (1985) *J. Exp. Med.*, **162**, 332–345.

Shimonkevitz,R., Colon,S., Kappler,J., Marrack,P. and Grey,H. (1984) *J. Immunol.*, **133**, 2067–2074.

Singh,B., Lee,K., Fraga,E., Wilkinson,A., Wong,M. and Barton,M. (1980) *J. Immunol.*, **124**, 1336–1343.

Spouge,J., Guy,H., Cornette,J., Margalit,H., Cease,K., Berzofsky,J. and DeLisi,C. (1987) *J. Immunol.*, **138**, 204–212.

Suzuki,G. and Schwartz,R. (1986) *J. Immunol.*, **136**, 230–239.

Taylor,P., Davey,J., Howland,K., Rothbard,J. and Askonas,B. (1987) *Immunogenetics*, **26**, 267–272.

Taylor,W.R. (1986) *J. Mol. Biol.*, **188**, 233–258.

Thomas,D., Hsieh,K., Schauster,J., Mudd,M. and Wilner,G. (1980) *J. Exp. Med.*, **152**, 620–628.

Thomas,D., Hsieh,K., Schauster,J. and Wilner,G. (1981) *J. Exp. Med.*, **153**, 583–587.

Thomas,J., Danho,W., Bullesbach,E., Fohles,J. and Rosenthal,A. (1981) *J. Immunol.*, **126**, 1095–1101.

Thornley-Lawson,D. and Israelson,E. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5384–5388.

Townsend,A., Rothbard,J., Gotch,F., Bahadut,B., Wraith,D. and McMichael,A. (1986) *Cell*, **44**, 959–968.

Watts,T., Gariepy,J., Schoolnik,G. and McConnell,H. (1986) *Proc. Natl. Acad. Sci. USA*, **82**, 5480–5484.

Zamvil,S., Mitchell,D., Moore,A., Kitamura,K., Steinman,L. and Rothbard,J. (1986) *Nature*, **324**, 258–260.

Received on October 19, 1987