Prediction of T cell Epitopes for HIV Vaccine Development by Computer-driven Algorithm

Anne S. De Groot, Bill M. Jesdale, Gabriel E. Meister, Neal I. Muni, and Caroline G. P. Roberts

TB/HIV Research Laboratory, International Health Institute,
Brown University School of Medicine, Providence RI, 02912

1. Overview of T cell epitope prediction algorithms

In the past 10 years, several computer-driven algorithms have been devised to take advantage of the alphabetic representation of protein sequence information to search for T cell epitopes. These algorithmssearch the amino acid sequence of a given protein for characteristics believed to be common to immunogenic peptides, locating regions that are likely to induce a cellular immune response *in vitro.* Given the rapid expansion of sequence data on geographic subtypes (clades) of HIV and individual HIV quasispecies, the application of these algorithms to HIV proteins may significantly reduce the number of regions which would require *in vitro* testing for immunogenicity, directing research to more promisingsegments ofHIVproteinsand thuspotentiallyreducing thetime andeffortneeded to develop HIV vaccines.

Computer-driven algorithms can identify regions of HIV proteins that contain epitopes and are less variable among geographic isolates; alternatively, computer-driven algorithms can rapidly identify regions of each geographic isolate's more variable proteins that should be included in a multi-clade vaccine. Furthermore, computer-driven searches can be weighted to reflect selected HLA alleles that aremostrepresentativeofgeographic populations orsubgroups withinone geographic area. Computer-driven searches can also be used as a preliminary tool to evaluate the evolution of immune response to an individual's own quasispecies. This text will review the development of computer algorithms for T cell epitope prediction, with a particular focus on the novel algorithm EpiMer, and will describe new directions for the application of computer algorithms such as EpiMer to HIV vaccine research.

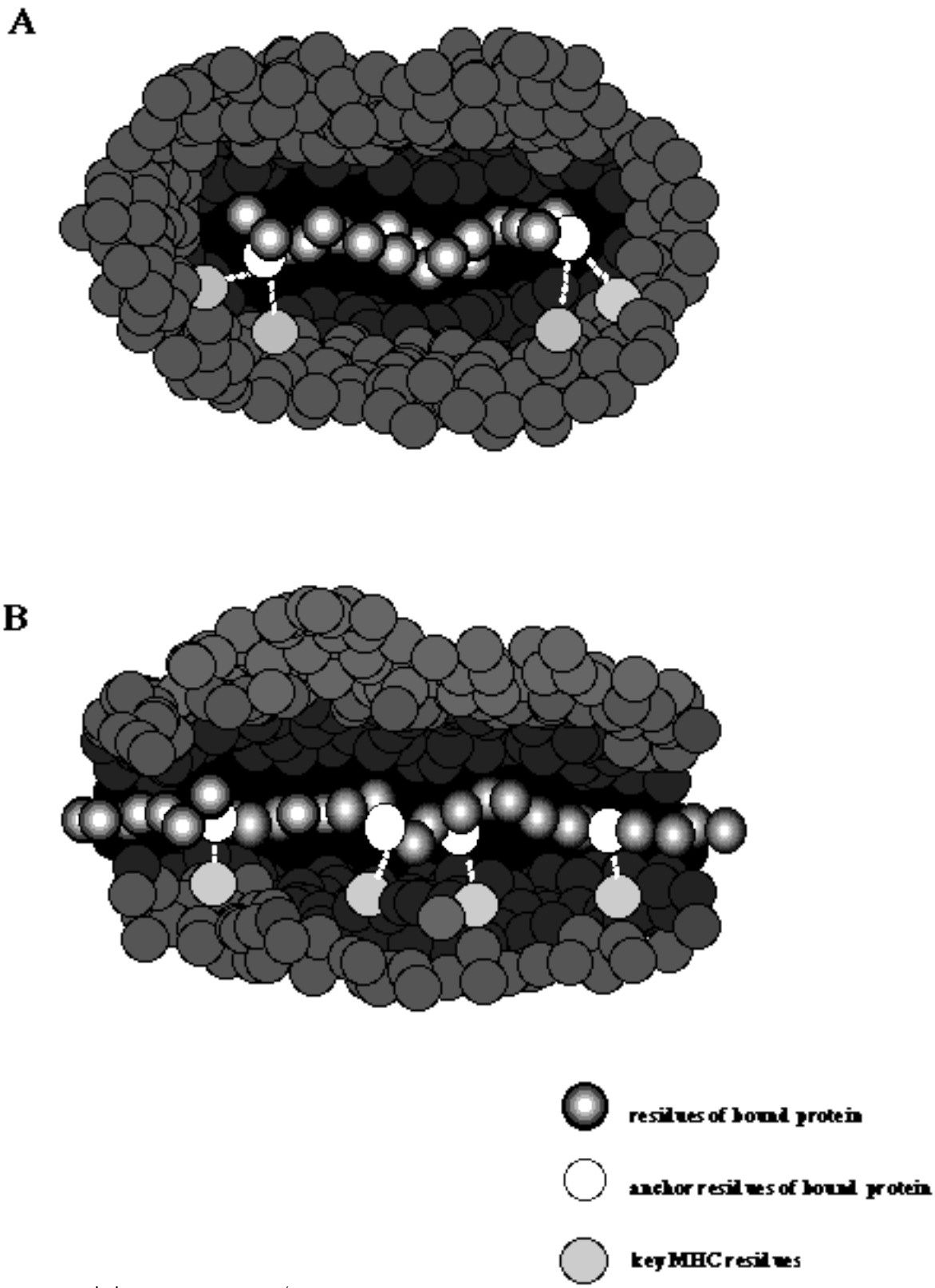
A. Characteristics of T cell epitopes

Peptides presented in conjunction with class I MHC molecules are derived from foreign or self protein antigens that have beensynthesized in thecytoplasm [1–3]. Peptides presented in the context of class II MHC molecules are usually derived from exogenous protein antigens [4–6]. Peptides binding to class I molecules are generally shorter (8–10 amino acid residues) than those that bind to class II molecules (8 to greater than 20 residues). An interpretation of peptides positioned in the binding cleft of class I and class II MHC molecules is shown in Figure 1.

The identification of T cell epitopes within protein antigens has traditionally been accomplished using a variety of methods, including the use of whole and fragmented native or recombinant antigenic protein, as well as the more commonly employed "overlapping peptide" method (Figure 2). The latter method for the identification of T cell epitopes within protein antigens involves the synthesis of overlapping peptideswhich span theentire sequence ofa given proteinantigen. These peptides are then tested for their capacity to stimulate T cell cytotoxic or proliferative responses *in vitro.*

As one might imagine, implementation of the overlapping peptide method is both cost- and labor-intensive. For example, to perform an assay using 15-amino acid-long peptides overlapping by 5 amino acids spanning a given antigen of length n (a small subset of all possible 15-mers spanning the protein), one would need to construct and assay $(n/5) - 1$ peptides. Yet, this method does not ensure the identification of all possible T cell epitopes, as potential sites can be "missed" between overlapping fragments.

The first research groups to suggest that computer algorithms based on patterns of amino acids might be used as a tool for discovering T cell epitopes were DeLisi and Berzofsky [7] and Rothbard and Taylor [8]. DeLisi and Berzofsky originally proposed the hypothesis that T cell antigenic peptides areamphipathicstructuresboundintheMHCgroove,withahydrophobicsidefacingtheMHCmolecule



binding motif." A library of such motifs has been generated for both Class I and Class II MHC alleles.


a. overlapping peptide method

andahydrophilicsideinteractingwiththeTcellreceptor[9]. RothbardandTaylor'salgorithmdescribes a similar periodicity for a smaller number of amino acid residues. The AMPHI algorithm, based on the DeLisi and Berzofsky observations and developed by Margalit et al.,[10] has been widely used for the prediction of T cell antigenic sites from sequence information alone.

Algorithms such as AMPHI, which are based on the periodicity of T cell epitopes, have been re-evaluated due to recent crystallographic determination of MHC structures with bound peptides. These peptidesweredemonstratedtobelyingextendedintheMHCgroove,innonalpha-helicalconformations [11,12]. An explanation of the predictive strength of AMPHI has been provided by Cornette et al.,[13] based on periodicity analysis of a table of motifs compiled by Meister et al., [14]. Essentially, AMPHI describes a common structural pattern of MHC-binding motifs, since the collection of MHC-binding motifs published to date appear to exhibit the same periodicity as an alpha helix. More recently, the rapid expansion of information on the nature of peptides that bind to MHC molecules has led to the evolution of a new class of computer-driven algorithms for vaccine development.

B. Algorithms based on MHC-binding motifs

MHC-bindingmotifsarepatterns ofaminoacidsthatappearto becommonto mostofthepeptides that bind to a specific MHC molecule. For example, a motif for a given MHC molecule might require thatalysine occuratpositionN+1(oneaminoacidfromtheaminoterminus), andavalineinpositionN + 8, while any amino acid is permitted at any of the other positions. In theory, such motifs may explain why MHC molecules are able to present many different peptides derived from different proteins, yet MHC restriction canstill occur. The peptide motif-MHC specificity appears to be due to the interaction of the amino acid side chains of certain conserved "anchor" residues (lysine and valine in the example given above) with pockets in the MHC peptide binding cleft (as diagrammed in Figure 1).

Identification of T cell epitopes by locating MHC-binding motifs in the sequence of a given protein has been shown to be effective when used to identify immunogenic epitopes for malaria [15] and for *Listeria monocytogenes*, [16] however, the number of regions of any given protein that contain single MHC motifs is usually much too large to be of any use for vaccine development. Furthermore, MHC-binding motifs appear to be relatively imprecise: only about one-third of peptides containing one of the current motifs that is said to predict binding to a given class I MHC allele have been shown to be bound by that MHC molecule, and in some cases, epitopes that do not contain known MHC-binding motifs have been described [17–19]. This may be due to missing information about the requirements for peptide-MHC interactions, or to errors in the descriptions of MHC-binding motifs in the literature. Inaddition,MHC-binding is necessarybutnot sufficientfor apeptide to be antigenic; thepeptide-MHC complex must still interact with the TCR of a neighboring cell, allowing the induction of a cellular immune response (reviewed in [20]).

Since 1992, members of the TB/HIV laboratory at Brown University have been developing a computer algorithm (EpiMer) that locates MHC-binding motifs in amino acid sequences of HIV proteins. In the process of developing this algorithm, we demonstrated that MHC-binding motifs tend to cluster within proteins [21]. Some of the clustering may be due to the similarity of certain MHC-binding motifs to one another, however, dissimilar motifs are also found to cluster. These motif-dense regions appear to correspond with peptides that may have the capacity to bind to a variety of MHC molecules (promiscuous or multi-determinant binders) and to stimulate an immune response in these various MHC contexts as well (promiscuous or multi-determinant epitopes).

EpiMer uses a library of MHC-binding motifs for class I and class II HLA alleles to predict antigenic sites within a protein that have the potential to induce an immune response in subjects with a varietyofgeneticbackgrounds. EpiMerlocatesmatchestoeachMHC-bindingmotifwithintheprimary sequenceof agiven protein antigen. Therelativedensity ofthese motifmatches isdetermined along the length of the antigen, resulting in the generation of a motif-density histogram. Finally, the algorithm identifies protein regions in this histogram with a motif match density above an algorithm-defined cutoff density value, and produces a list of subsequences representing these clustered, or motif-rich regions (Figure 3). The regions selected by EpiMer may be more likely to act as multi-determinant binding peptides than randomly chosen peptides from the same antigen, due to their concentration of MHC-binding motif matches. An example of a multi-determinant epitope is shown in Figure 4.

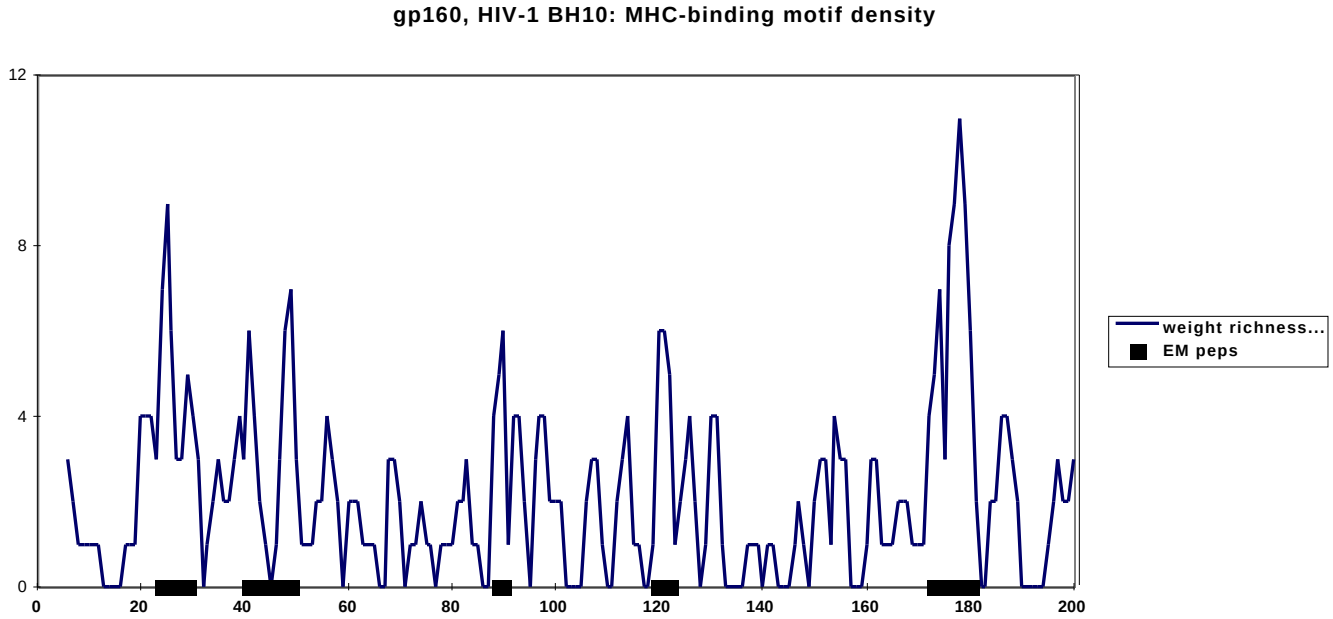


Figure 3. MHC-binding motif clustering for gp160 of strain BH 10 A histogram of the density of MHC-binding motif matches along the sequence of the gp160 protein of HIV BH 10 is shown here, to illustrate the EpiMer method of putative epitope identification. For this analysis, both class I and II MHC-binding motifs were used in our search. Peptides that include peaks of motif density, such as the 10- to 25-mers including amino acids 19 to 34 (14 motifs), 36 to 54 (14 motifs), 84 to 95 (6 motifs), 115 to 127 (7 motifs) and 168 to 185 (22 motifs) shown in this example, are predicted as putative T cell epitopes by the EpiMer algorithm. The EpiMer peptides are shown in bold, and are slightly shorter than the stated predictions because the midpoint of the amino terminal 11-mer reading frame ofthe predicted peptide to themidpoint ofthe carboxy terminal 11-mer reading frame are designated in this picture, rather than the full length of the predicted peptide.

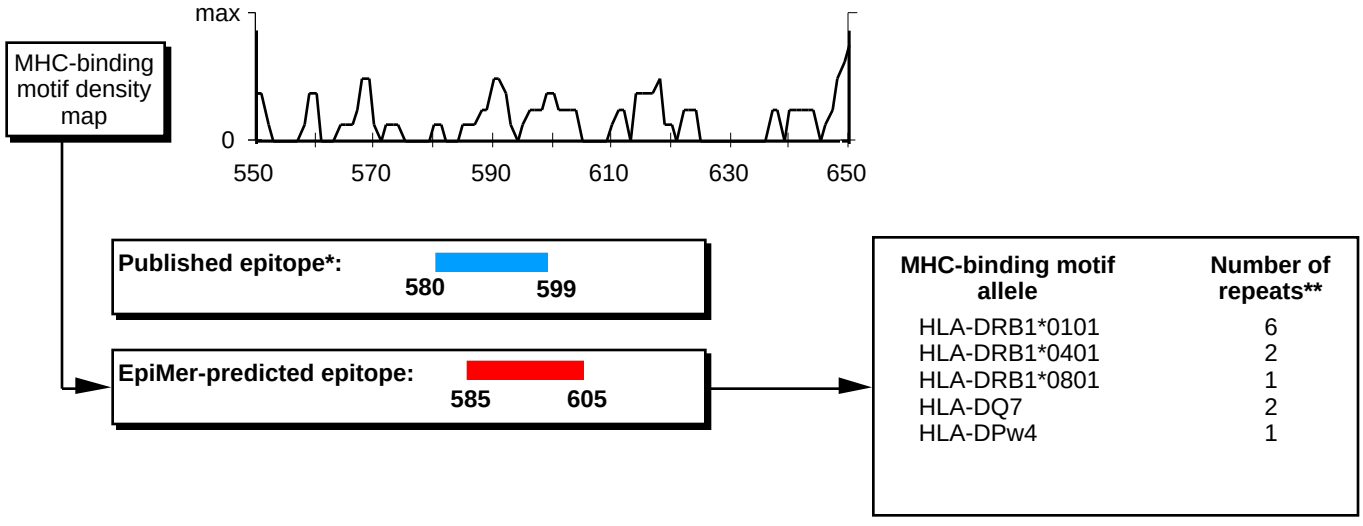


Figure 4. Multi-determinant peptide. An example of a multi-determinant peptide is shown. The MHC-binding motifs for the predicted peptide are shown at the right. Note that HLA DRB1*0101 motif occurs six times, that five unique motifs can be identified, and that a total of 12 potential MHC-binding regions are contained within this protein sequence.

The MHC-binding motif library used by EpiMer for its searches is updated regularly from the literature. This list can be tailored for a number of different types of searches. For example, one can use the entire MHC-binding motif library to identify peptides that contain both MHC class I and class II binding motifs; one can restrict the list of binding motifs used in the searches to class I or class II, and one can tailor the search to the set of MHC alleles of geographic subpopulation or even those of a single individual.

The utility ofcomputer-algorithm driven predictions for *in vitro* and *in vivo* research wasrecently demonstrated in an analysis of peptides predicted by the EpiMer algorithm from *Mycobacterium tuberculosis* (Mtb) protein sequences. Twenty-seven of 28 EpiMer peptides derived from Mtb proteins stimulated immune responses (proliferation) in peripheral blood cells from Mtb immune subjects [21]. There was a good correlation between the number of motifs per peptide and the number of responders to the peptide in a population of Mtb-infected individuals ($p < 0.001$), and 40 percent of the variation in the relationship between the motifs and the responses could be explained by the presence or absence of MHC-binding motifs [22]. As only about a third of peptides that are predicted using single MHC-binding motifs are shown to bind and to stimulate immune responses, the relationship between the EpiMer predictions and the number of responders to the peptides was much better than might have been expected. We believe that the selection of regions that are MHC-binding motif-dense increases the likelihood that the predicted peptide contains a "valid" motif, and furthermore, that the reiteration of identical motifs may contribute to peptide binding [23].

Additional MHC-binding motif-based algorithms have been described by Parker et al. [24], and Altuvia et al, [25]. In these algorithms, binding to a given MHC molecule is predicted by a linear function of the residues at each position, based on empirically defined parameters, and in the case of Altuvia et al., known crystallographic structures are also taken into consideration [25,26]. DeLisi et al. have proposed an alternative method of determining MHC-binding peptides, based on the free energy relationshipsof each amino acid in the predicted peptide, and using thisinformation to analyze whether the tertiary structure of the peptide conforms to a predetermined MHC-binding peptide configuration [27,28]. Brusic and colleagues are using artificial neural networks to determine the "rules" for binding to MHC molecules from the complete list of binding peptides that have been published for each of the human HLA alleles [29]. Hammer et al. [30] describe a technique known as "peptide side chain scanning", which they used to predict binding peptides for the MHC allele DRB1*0401. This allowed the construction of a matrix of all possible amino acid side chain effects for a single MHC-binding motif, whichwaslater converted into an algorithm able to run through a protein's primary structure and predict, within reasonable error, the binding capacities of all possible peptides of a fixed length to a singleMHCmolecule. NoneofthesealgorithmS, withtheexceptionofEpiMer,havebeentested *invivo*. Should any of these computer-driven variations on "motif matching" prove to be accurate predictors of peptides that bind to individual MHC alleles, they may be easily incorporated as subprograms into the motif-library portion of EpiMer, and might improve the algorithm's overall predictive capacity.

Most of the computer-driven algorithms described in this text depend on published information on MHC-binding motifs, or on knowledge of the crystallographic structure of peptides within MHC binding grooves. One methodological concern when designing a multiple binding motif-based predictive algorithm is theaccuracy of the MHC-bindingmotifs used to predictputativeepitopes, andthus the overall validity of the motif database. Previously reported motifs are often redefined in the literature, afterpeptidetruncationandalaninesubstitutionexperimentsareperformed; likewise,newemphasishas been placed on the role of protein processing and on the identification of specific amino acid residues at non-anchor sites, which interfere with the relative capacities of peptides to bind to the MHC cleft [31,32]. In addition, several MHC-binding motif databases have been constructed. Rammensee et al. [33] have published a motif database, aided by the alignment of actual MHC-binding peptides and known T cell epitopes. A new prediction algorithm based on the Rammensee motifs has been developed in the TB/HIV Research Laboratory (Bill M. Jesdale and Gabriel E. Meister, unpublished data). Brusicet al. [34]have takenthis MHC motiflibrary concept furtherby providing an Internet-accessible database of binding motifs and peptides known to bind with affinity to MHC molecules.

An important consideration when comparing the different computer-driven models described above is that these methods for epitope prediction are not mutually exclusive. As the contributions of

sidechainsandtertiarypeptide structureto peptide-MHCbindingarebetterquantified, thedevelopment of a computer algorithm that predicts T cell epitopes based on a matrix of side chain information such as the matrix described by Hammer[35] will only be a matter of time. The identification of novel structuralfeatureswhichareableto independentlypredictpeptide bindingorimmunogenicity, andtheir subsequent synthesis into a combined algorithm with statistically verifiable predictive capacity, may allowadramaticreductioninthe timeandeffortrequired tosynthesizeandtestpotentialTcellantigenic sitesforHIVproteins, byallowingthepredictionofsiteswithahighconcentrationofantigenicfeatures.

2. Applications of T cell epitope algorithms to HIV research

A. Searching for T cell epitopes

Identification of T cell epitopes that stimulate cell-mediated immunity is essential to the development of an HIV vaccine. The identification of HIV peptide epitopes that contain clusters of MHC-binding motifs representing multiple HLA alleles from HIV protein sequences may be useful for HIV vaccine development.

There appear to be more stringent binding criteria for class I-restricted binding peptides, and few multi-determinantclass Iepitopeshavebeenidentified forany pathogen. However, several HIV protein regions that contain multiple overlapping class-II restricted epitopes, also known as "promiscuous" or multi-determinant peptides, have been identified in mice andhumans. Such regionsmight beimportant to include in thesynthesis ofmultiple antigenic peptides (MAPS) for HIV vaccine development, particularly ifa multi-determinant Tcell epitope is required forboosting immune response to B cell epitopes.

The EpiMer algorithm isreadily applied to HIV protein sequences. The efficiency and sensitivity of the EpiMer algorithm for detecting published T cell epitopes was recently compared to that of the overlapping method, for the HIV proteins, gp160, nef, tat, and gag (Roberts et al., manuscript submitted [36]). The EpiMer algorithm predicted putative T cell epitopes from protein sequences for HIV-1nef, gp160, gag p55, andtat that required fewerpeptides and thereforefewer amino acidresidues to be synthesized than either AMPHI-predicted peptides or overlapping peptides. For the four HIV-1 proteins, EpiMer predicted 43 peptide epitopes, AMPHI predicted 68 peptides , and the overlapping peptide method (20 amino acid long peptides overlapping by 10 amino acids) would have required 161 peptides. When the number of published epitopes that correlated with predictions was evaluated in terms of the number of amino acids synthesized using each method of prediction (SAA), the EpiMer methodofpredictionwas2.4foldbetterthantheoverlappingmethod. Asummaryofthesecomparisons is shown in Table 1.

Table 1 Efficiency and sensitivityof the Overlapping method, compared toEpiMer, for the HIV proteins nef, gag, gp160, and tat.

Method	**Overlapping**	**EpiMer**
Percent efficiency	60%	62%
Range	43%–100%	61%–64%
Percent sensitivity	100%	59%
Range	100%–100%	22%–86%
Average sensitivity per amino acid	2.7	4.9
Range	0.6–6.4	1.3–8.1
Average Δ sensitivity per AA	(ref)	2.4

Efficiency, = (total length, in amino acid residues, of the peptides that overlap by at least eight or eleven amino acid residues with Class I or Class II published epitopes, respectively)/(total length, in amino acidresidues, ofall putativeepitopes identified bythe algorithm in question). Sensitivity, S,(number of publishedepitopesthatwerepredictedbythealgorithminquestion)/(total numberofpublishedepitopes for the protein). Sensitivity per amino acid (SAA), = 1000 $\times$ x (Sensitivity)/(total length, in amino acids, of peptides to be synthesized). Δ Sensitivity/AA, (ΔSAA) = (Sensitivity per amino acid residue for a given method)/(Sensitivity per amino acid residue of the overlapping peptide method).

A complete list of the MHC-binding motifs contained within the most widely used laboratory strains of HIV-1 proteins, based on the MHC-binding motif list compiled by the TB/HIV Research Laboratory, is being developed for their Web Site (http://www.brown.edu/Research/TB-HIV-Lab/); a partial list for amino acids 628 to 678 of the HIV-1 BH 10 protein gp160 is shown in Table 2. As demonstrated, regions of HIV proteins that contain as many as 20 to 30 MHC-binding motifs can be identifiedusingEpiMer. Suchregionsshouldbegoodcandidatesforinclusion inasubunitHIVvaccine.

ApplicationofMHC-bindingmotifsto HIVvaccinedevelopment mayberestrictedbytheamount of sequence variation in individual quasispecies, HIV strains, and HIV clades, as well as by the MHC background of the target populations. One might consider evaluating regions of MHC clustering that occur in sites of low HIV sequence variability, as shown in Figure 5. The region 130 to 160, which has a great deal of inter-strain variation described by the variability plot, might best be avoided for subunit HIV vaccine development. HIV peptide epitopes which contain multiple MHC-binding motifs, either conserved across HIV strains or derived from several different HIV strains, may be ideal candidates for inclusion in a multisubunit vaccine.

An alternative to searching for conserved regions of HIV proteins would be to identify regions of the sequences that predominate in the clades that are most likely to be presented in the context of the MHC molecules of the geographic subpopulations of interest. If MHC-binding motif-based peptides are to be used in subunit vaccines, the best strategy may be to custom design the peptides, using the sequences of the HIV clades that are prevalent in that populations and the set of MHC alleles that are alsoprevalent in that population. We have proposed one method ofweighting predicted peptides by the prevalence of MHC-binding motifs [37]; the DeLisi laboratory has proposed yet another method [28].

B. Evaluating the effect of immune response on the evolution of HIV

An additional application of EpiMer might be to evaluate the effect of pressure from the immune system of the individual on the HIV quasispecies of that individual. Ongoing research has suggested that rapid progression might be related to the capability of the virus to avoid immune detection through variation at the MHC-binding anchorsof agivenT cellepitope, orthrough variation atthe TCR binding site. To date, several laboratories have described *in vitro* evidence for escape mutations in the epitope of a given individual [38,39]. We have examined the evolution of class I MHC-binding peptides in HIV-1 quasispecies in the contexts of clinically quiescent HIV-1 infection and rapid progression to advanced disease, by implementing EpiMer to predict MHC-binding peptides from primary protein sequences [40]. Using each patient's own MHC allele subset to tailor the EpiMer searches, regions of the patient's own quasispecies can be searched for putative MHC-binding peptides. This search can be repeated for quasispecies isolated from the patient at each of several timepoints, and analyzed for patterns of MHC-binding motif escape, or replacement by an alternate binding region. This novel approach identifies regions of HIV quasispecies that should be the focus of binding assays and epitope mapping which may improve our comprehension of host immune response to HIV.

Summary

Identification ofT cellepitopes that stimulate cell-mediated immunity isessential to HIV vaccine development. Computer driven algorithms for T cell epitope prediction appear to provide rapid and relatively inexpensive means of T cell identification for *in vitro* investigations. The EpiMer algorithm, described in this text and in more detail in reference [14], identifies peptide epitopes from HIV proteins by identifying clustering of MHC-binding motifs within the protein sequences. Peptide epitopes containingmultipleMHC-bindingmotifsmaybeimmunogenicinindividualsfromavarietyofgeneticbackgrounds. Identification ofsuchclustersmay improve theimmunogenicityofagivenpeptide, andpermit the development of a subunit vaccine that can induce immunity to multiple strains and clades of HIV.

Identification of T cell epitopes within the sequences from quasispecies of HIV-infected individuals may also permit the investigations of the evolution of HIV in response to host immune pressure. While the relationship between MHC-binding motifs and immunogenicity is less than absolute, the utilization of computer driven algorithms such as EpiMer may permit the identification of regions of increased interest for *in vitro* confirmation of HIV evolution within an individual or within a given geographic subpopulation.

Amino Acids			Amino Acid Sequence	Motif Match
***628**	**-**	**647**	**WMEWDREINNYTSLIHSLIE**	
629	-	638	MEWDREINNY	HLA-B*44
629	-	638	MEWDREINNY	HLA-DPw4
629	-	638	MEWDREINNY	HLA-DRB1*0301
629	-	638	MEWDREINNY	HLA-DRB1*0801
630	-	638	EWDREINNY	HLA-A1
631	-	639	WDREINNYT	HLA-DRB1*0401(DR4Dw4)
633	-	641	REINNYTSL	HLA-B*40012
633	-	641	REINNYTSL	HLA-B40
633	-	641	REINNYTSL	HLA-B44
633	-	641	REINNYTSL	HLA-Cw*0301
635	-	644	INNYTSLIHS	HLA-DRB1*1501
637	-	645	NYTSLIHSL	HLA-A24
637	-	645	NYTSLIHSL	HLA-Cw*0401
637	-	645	NYTSLIHSL	HLA-Cw*0602
637	-	645	NYTSLIHSL	HLA-Cw*0702
637	-	645	NYTSLIHSL	HLA-DQ3.1
638	-	646	YTSLIHSLI	HLA-DQ3.1
***678**	**-**	**712**	**WLWYIKLFIMIVGGLVGLRIVFAVLSVVNRVRQGY**	
679	-	687	LWYIKLFIM	HLA-DR1
679	-	688	LWYIKLFIMI	HLA-DPw4
679	-	688	LWYIKLFIMI	HLA-DRB1*0801
679	-	688	LWYIKLFIMI	HLA-DRB1*1501
680	-	688	WYIKLFIMI	HLA-A24
680	-	688	WYIKLFIMI	HLA-Cw*0301
680	-	688	WYIKLFIMI	HLA-DPA1*0102/DPB1*0201
681	-	689	YIKLFIMIV	HLA-Cw*0602
681	-	689	YIKLFIMIV	HLA-DPA1*0102/DPB1*0201
681	-	689	YIKLFIMIV	HLA-DR1
681	-	689	YIKLFIMIV	HLA-DRB1*0401(DR4Dw4)
682	-	690	IKLFIMIVG	HLA-DQ7
682	-	690	IKLFIMIVG	HLA-DRB1*0401(DR4Dw4)
682	-	691	IKLFIMIVGG	HLA-DRB1*1501
684	-	692	LFIMIVGGL	HLA-Cw*0401
684	-	692	LFIMIVGGL	HLA-Cw*0602
684	-	692	LFIMIVGGL	HLA-DR1
685	-	693	FIMIVGGLV	HLA-DRB1*0101
686	-	695	IMIVGGLVGL	HLA-DPw4
686	-	695	IMIVGGLVGL	HLA-DRB1*1501
687	-	695	MIVGGLVGL	HLA-A*0205
687	-	695	MIVGGLVGL	HLA-DR1
688	-	696	IVGGLVGLR	HLA-A*3302
688	-	696	IVGGLVGLR	HLA-DQ3.1
688	-	697	IVGGLVGLRI	HLA-A68
689	-	697	VGGLVGLRI	HLA-B*5101
689	-	697	VGGLVGLRI	HLA-B*5102
689	-	697	VGGLVGLRI	HLA-B*5103
689	-	697	VGGLVGLRI	HLA-DQ3.1
689	-	697	VGGLVGLRI	HLA-DQ7
689	-	697	VGGLVGLRI	HLA-DRB1*0101
689	-	698	VGGLVGLRIV	HLA-DPw4
690	-	698	GGLVGLRIV	HLA-B*5102
690	-	698	GGLVGLRIV	HLA-B*5103
691	-	699	GLVGLRIVF	HLA-A3
691	-	699	GLVGLRIVF	HLA-B*1501
692	-	700	LVGLRIVFA	HLA-DR1
692	-	700	LVGLRIVFA	HLA-DRB1*0401(DR4Dw4)
692	-	701	LVGLRIVFAV	HLA-DPw4
692	-	701	LVGLRIVFAV	HLA-DRB1*0801
693	-	700	VGLRIVFA	HLA-B*7801
693	-	701	VGLRIVFAV	HLA-B*5102
693	-	701	VGLRIVFAV	HLA-B*5103

Table 2. Amino acid sequences of the EpiMer-predicted epitopes for the amino acid residues 628 to 678 of the gp160 protein are listed (in boldface), as are the individual MHC-binding motif matches found within each peptide. These two predicted epitopes overlap with published epitopes for this same HIV-1 strain.

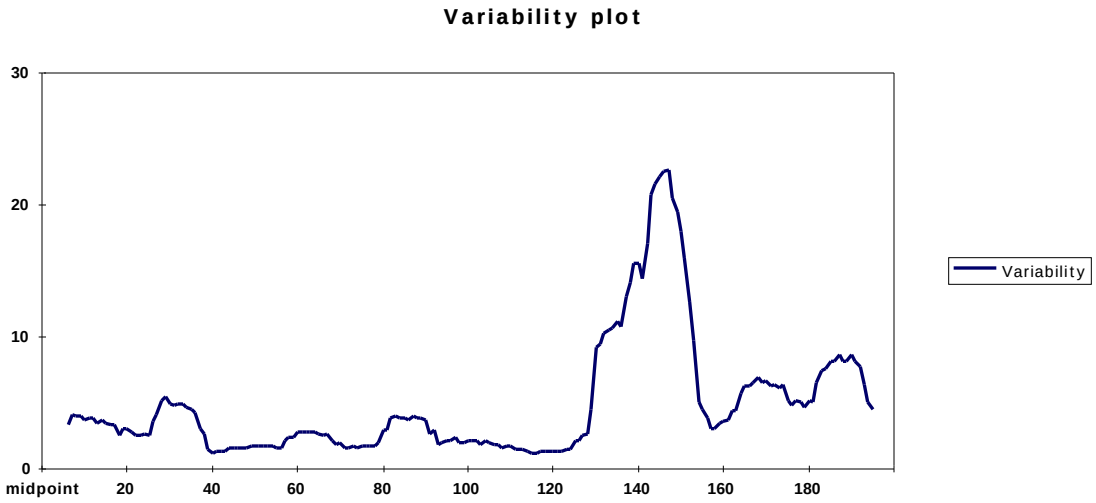


Figure 5. gp160 (variability plot) The mean variability of the 11-residue segments of known gp160 sequences (Los Alamos HIV Sequence Database) are shown aswell. Variability = (number of different amino acids at a given position)/(frequency of the most common amino acid at that position).

References

[1] Matsumura, M, Fremont, DH, Peterson, PA and Wilson, IA. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* 1992; 257:927–934.

[2] Germain, RN and Margulies, DH. The biochemistry and cell biology of antigen processing and presentation. *Annu Rev Immunol* 1993; 11:403–450.

[3] Falk, K, Rötzschke, O, Stevanovic, S, Jung, G and Rammensee, H-G. Allele-specific motifs revealed by sequencing ofself-peptideseluted fromMHC molecules. *Nature* 1991; 351:290–296.

[4] Chicz, RM, Urban, RG, Gorga, JC, Vignali, DAA, Lane, WS and Strominger, JL. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. *J Exp Med* 1993; 178:27–47.

[5] Srinivasan, M, Domanico, SZ, Kaumaya, PTP and Pierce, SK. Peptides of 23 residues or greater are required to stimulate a high affinity class II-restricted T cell response. *Eur J Immunol* 1993; 23:1011–1016.

[6] Rötzschke, O and Falk, K. Origin, structure and motifs of naturally processed MHC class II ligands. *Curr Opin Immunol* 1994; 6:45–51.

[7] DeLisi, C and Berzofsky, JA. T-cell antigenic sites tend to be amphipathic structures. *Proc Natl Acad Sci USA* 1985; 82:7048–7052.

[8] Rothbard, JB and Taylor, WR. A sequence pattern common to T cell epitopes. *EMBO J* 1988; 7:93–100.

[9] Spouge, JL, Guy, HR, Cornette, JL, Margalit, H, Cease, K, Berzofsky, JA, and DeLisi, C. Strong conformational propensities enhance T cell antigenicity. *J Immunol* 1987; 138:204–212.

[10] Margalit, H, Spouge, JL, Cornette, JL, Cease, KB, DeLisi, C and Berzofsky, JA. Prediction of immunodominant helper T cell antigenic sites from the primary sequence. *J Immunol* 1987; 138:2213–2229.

[11] Cornette, JL, Margalit, H, DeLisi, C, and Berzofsky, JA in *The Amphipathic Helix*, R. M. Epand, Ed. CRC Press, Boca Raton, FL, 1993, 333–345.

[12] Stern, LJ, Brown, JH, Jardetzky, TS, Gorga, JC, Urban, RG, Strominger, JL, ,Wiley, DC. *Nature* 1994; 368:215–221.

[13] Cornette, JL, Margalit, H, DeLisi, C, and Berzofsky, JA. *Proc Nat Acad Sci USA* 1995, 92:8368–3872.

[14] Meister, GE, Roberts, CGP, Berzofsky, JA, De Groot, AS. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from *Mycobacterium tuberculosis* and HIV protein sequences. *Vaccine 1995*; 13:581–591.

[15] Hill, AVS, Elvin, J, Willis, AC, Aidoo, M, Allsopp, FM, Gotch, XM, Gao, T, Takiguchi, M, Greenwood, BM, Townsend, ARM, et al., Molecular analysis of the association of HLA-B53 and resistance to severe malaria *Nature* 1992; 340:434–439.

[16] Pamer, EG, Harty, JT, Bevan, MJ, Precise prediction ofa dominantclass IMHC-restricted eptiope of *Listeria monocytogenes Nature* 1991; 353:852.

[17] Lipford, GB, Hoffman, M, Wagner, H, and Heeg, K. Primary *in vivo* responses to ovalbumin. Probing the predictive value of the Kb binding motif. *J Immunol* 1993; 150:1212–1222.

[18] Nijman, HW, Houbiers, JGA, Vierboom, MPM, van der Burg, SH, Drijfhout, JW, D'Amaro, J et al. Identification of peptide sequences that potentially trigger HLA-A2.1-restricted cytotoxic T lymphocytes. *Eur J Immunol* 1993; 23:1215–1219.

[19] Calin-Laurens V, Trescol-Biémont M-C, Gerlier D and Rabourdin-Combe, C. Can one predict antigenic peptides for MHC class I-restricted cytotoxic T lymphocytes useful for vaccination? *Vaccine 1993*; 11:974–978.

[20] Sercarz, E, Lehmann, P, Ametani, A, Benichou, G, Miller, A, and Moudgil, K. Dominance and crypticity of T cell antigenic determinants. *Annu Rev Immunol* 1993; 11:729–766.

[21] Meister, GE, Roberts, CGP, Berzofsky, JA, De Groot, AS. *Mycobacterium tuberculosis* peptide epitopespredictedbytwonovelepitopeidentificationalgorithms, *Vaccines95*,ColdSpringHarbor Laboratory, Cold Spring Harbor NY, 1995.

[22] De Groot AS, Carter EJ, Roberts CGP, Edelson BT, Jesdale BM, Meister GE, Houghten RA, Montoya, J, Romulo RC, Berzofsky JA, Ramirez BDLL. A novel algorithm for the efficient identification of T cell epitopes: prediction and testing of candidate TB vaccine peptides in genetically diverse populations, to be published in *Vaccines 96* Cold Spring Harbor Laboratory, Cold Spring Harbor NY 1995.

[23] Sette, A., Sidney, J., Albertson, M., Miles, C., Colon, S.M., Pedrazzini, T., Lamont, A.G., and Grey, H.M. A novel approach to the generation of high affinity class II-binding peptides. *J. Immunol* 1990, 145, 1809–1813.

[24] Parker, KC, Bednarek, MA, and Coligan, JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side chains. *J. Immunol* 1994, 152:163–175.

[25] Altuvia, Y, Schueler, O, Margalit, H, Ranking potential binding peptides to MHC molecules by a computational threading approach, *J Mol Biol* 1995; 249:244–250.

[26] Altuvia, Y, Berzofsky, JA, Rosenfeld, R, and Margalit, H. Sequence features that correlate with MHC restriction. *Molecular Immunology 1994*, 31:1–19.

[27] S. Vajda S and DeLisi C, Determining Minimum Energy Conformations of Polypeptides by Discrete Dynamic Programming. *Biopolymers* 1990; 29:1755–1772.

[28] Gulukota, K, Vajda, S, and DeLisi, C. Peptide Docking using dynamic programming, *J Comp Biol*, submitted.

[29] Brusic, V, Rudy, G, Harrison LC. Prediction of MHC-binding peptides using artificial neural networks, In: Stonier RJ, Yu XS, eds., *Complex systems, mechanisms of adaption*. Amsterdam; IOS Press 1994:253–260.

[30] Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, and Sinigaglia F. Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning. *J Exp Med* 180; 1994:2353–2358.

[31] Boehncke, W-H, Takeshita, T, Pendleton, CD, Houghten, RA, Sadegh-Nasseri, S et al. The importance of dominant negative effects of amino acid side chain substitution in peptide-MHC molecule interactions and T cell recognition.*J Immunol* 1993; 150:331–341.

[32] Dicke, LR, Aldrich, C, Jameson, CR, Moomaq, BC, Praminic, CK, Coyle, GN, DeMartino, MJ, Bevan, JM. Proteolytic processing of ovalbumin and B-galactosidase by the proteasome to yield antigenic peptides. *J Immunol* 1994, 152:3884–3894.

[33] Rammensee, H-G, Friede, T and Stevanovic, MHC ligands and peptide motifs: first listing. S. *Immunogenetics* 1995; 41:178–228.

[34] Brusic, V, Rudy, G, Harrison, LC. MHCPEP: a database of MHC-binding peptides. *Nuc Acids Res* 1994, 22, 3663–3665.

[35] Hammer J., New methods to predict MHC Binding sequences within protein antigens, *Current Opinion in Immunology* 1995, 7:263–269.

[36] Roberts, CGP, Meister, GE, Jesdale, BT, Lieberman, J, Berzofsky, J, and AS De Groot, Identification of HIV peptide epitopes by a novel algorithm, submitted 1995.

[37] De Groot et al., Predicting Mtb epitopes for geographic subpopulations: progress on the development of a novel TB vaccine for West Africans. mansucript in preparation.

[38] Philips RE, McMichael, AJ,How does the HIV escape cytotoxic Tcell immunity, *Chem Immunol* 1993. 56:150–164.

[39] Safrit, JT, Lee, AY, Andrews, CA, Koup RA. A region of the third variable loop of HIV-1 gp120 is recognized by HLA B7 restricted CTLs from two acute seroconversion patients, *J Immunol*. 1994, 153:3822–3830.

[40] Jesdale, BM,Santiago,MLO,Wolinsky, SM,Korber,BT,Koup,RA,andDeGroot, AS,Evolution of HIV-1 quasispecies: sequence diversity in predicted class I MHC-binding regions over time (env C2-V5), manuscript in preparation.