Physicochemical correlation between amino acid sites in short sequences under selective pressure

D. S. Campo[1], Z. Dimitrova[1] and Y. Khudyakov[1]

[1]Molecular Epidemiology & Bioinformatics Laboratory, Division of Viral Hepatitis, Centers for Disease Control and Prevention. 1600 Clifton Rd, Atlanta, GA 30333.

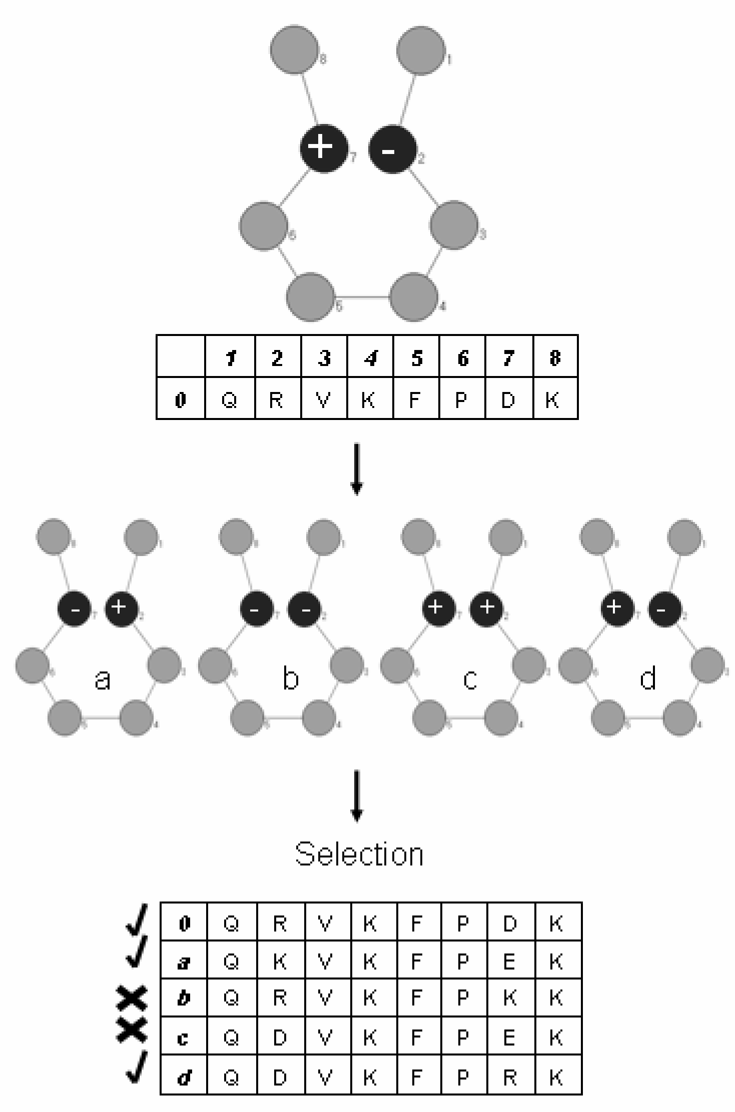
Abstract. The activities and properties of proteins are the result of interactions among their constitutive amino acids. In the course of natural selection, substitutions which tend to destabilize a particular structure may be compensated by other substitutions which confer stability to that structure. Patterns of coordinated substitutions were studied in two sets of selected peptides. The first is a set of 181 amino acid sequences that were selected *in vitro* to bind a MHC class I molecule (K^b). The second is a set of 114 sequences of the Hypervariable Region 1 of Hepatitis C virus, which, originating from infected patients, result from natural selection *in vivo*. The patterns of coordinated substitutions in both datasets showed many significant structural and functional links between pairs of positions and conservation of specific selected physicochemical properties.

Keywords: physicochemical properties, amino acid, covariation, selection.

1 Introduction

Experimental and quantitative analyses of proteins often assume that the protein sites are independent, i.e., the presence of a residue at one site is independent of residues at other sites. However, the activities and properties of proteins are the result of interactions among their constitutive amino acids (aa) and, therefore, substitutions which tend to destabilize a particular structure and/or function are probably compensated by other substitutions that confer stability [1]. For example, if a salt bond were important, a substitution of the positively-charged residue with a neutral residue would need to be compensated by a nearby residue substituting from a neutral to a positive residue (Fig. 1). Similarly, a substitution involving a reduction of volume in the protein core might cause a destabilizing pocket which only one or a few adjacent residues would be capable of filling. Sites which are structurally or functionally linked will tend to evolve in a correlated fashion due to the compensation process [1].There is experimental evidence indicating that proteins contain pairs of covariant sites, identified both by analysis of families of natural proteins with known structures [2-7] and by site-directed mutagenesis whereby individual changes are introduced in proteins [8-10].

Independent mutations among functionally-linked sites would be disadvantageous, but simultaneous or sequential compensating mutations may allow the protein to retain function [11]. Furthermore, there are constraints on aa replacements that arise for functional reasons, such as aa bias at recognition sites related to DNA binding in transcriptional regulators. Evolutionarily-related sequences should contain the vestiges of these effects in the form of covariant pairs of sites [12] and these interactions can be manifested in covariation between substitutions at pairs of alignment positions in a multiple sequence alignment. The analysis of covariation has been used in protein engineering [13], sequence-function correlations [14, 15], protein structure prediction [5, 12, 16-26] and in finding important motifs in viral proteins [27-30]. Recent analyses confirmed that highly coordinated sites are often functionally related and/or spatially coupled, with coevolving positions being frequently located in regions critical for protein function, such as active sites and surfaces involved in molecular interactions with other proteins [14, 31-35].



	1	2	3	4	5	6	7	8
0	Q	R	V	K	F	P	D	K

✓	0	Q	R	V	K	F	P	D	K
✓	a	Q	K	V	K	F	P	E	K
✗	b	Q	R	V	K	F	P	K	K
✗	c	Q	D	V	K	F	P	E	K
✓	d	Q	D	V	K	F	P	R	K

Fig. 1. Schematic representation of coordinated substitutions in a pair of aa sites forming a salt bond in a protein domain. Sequences that contain residues of the same charge at positions 2 and 7 are unstable (b and c) and are eliminated during natural selection. Sequences containing residues of different charges that are stable (o, a and d) can occur in a multiple sequence alignment.

In this paper, the patterns of coordinated substitutions were studied in two sets of selected peptides. The first is a set of 181 aa sequences that had been selected *in vitro* to bind a MHC class I molecule (K^b) [36, 37]. The second is a set of 114 sequences from Hypervariable Region 1 (HVR1) of Hepatitis C virus (HCV) generated *in vivo* from infected patients, which was used to understand the effects of natural selection on the pattern of coordinated substitutions. In both cases the process of selection over the structure and/or function of the peptide constrained the sequence variability and we found vestiges of these effects in the form of covariant pairs of sites.

2 Methods

2.1 Datasets

A previously published dataset of 310 peptides [36, 37] was used to investigate the effects of *in vitro* selection on the pattern of coordinated substitutions. This dataset gives the peptide aa sequence and corresponding binding to a MHC class I molecule (K^b) as a binary (yes/no) outcome. The complete dataset has 310 such observations (181 binders and 129 non-binders) and was originally obtained by random sampling from a large ($>10^7$) library of peptides, so there is no evolutionary history linking the peptides. The binding between two proteins generally involves short-range non-covalent interactions based on electrostatic charge, hydrogen bonding and van der Waals interactions. The specificity of the binding depends on the physicochemical properties of the constituent aa residues of both molecules and, therefore, the binding to the MHC class I molecule must select the conservation of some physicochemical properties in this subset of aa sequences. In this paper we wanted to know if the selection for this known function (binding) could be detected in the form of physicochemical correlation between aa sites.

HCV is a major cause of liver disease worldwide. The global prevalence of HCV infection is estimated to be 2.2%, representing 130 million people [38]. HCV causes chronic infection in 70-85% of infected adults [39]. There is no vaccine against HCV and current anti-viral therapy is relatively toxic, being effective in 50–60% of patients treated [40]. HCV is a single-stranded RNA virus of approximately 9400 nucleotides belonging to the *Flaviviridae* family [41]. The HVR1, located between positions 384 to 410 of the structural E2 protein, is the most intensively studied part of the HCV genome. However, the understanding of its function remains very limited [42] and it is not clear whether its high genetic heterogeneity is an immunological decoy or is related to a biologically relevant function [43]. Here, we studied the sequence variability of HVR1 in order to establish physicochemical correlations between aa sites, which could be due to the pressure of an unknown function that selected the conservation of physicochemical properties.

2.2 Sequences and alignment

The previously published dataset of 310 peptides [36, 37] was obtained from the following web address: http://newfish.mbl.edu/Lab/Resources. All these peptides were created in the same phage display library and had the same length so they were easily aligned. Two hundred and eight complete genome HCV sequences were obtained from “The Los Alamos HCV Sequence Database” [44] during early 2006. Of these 208 sequences, the following were excluded: recombinants, chimeras, patents, non-human hosts, a genotype other than 1b, and epidemiologically related sequences. This process left 114 different HCV 1b complete genome sequences, which were aligned using ClustalW [45]. The viral protein H77 (GenBank Accession Number NC_004102) was used as a reference sequence throughout this study.

2.3 Physicochemical properties of aa

A study by Chelvanayagam et al. [31] found that the analysis of covariation involving different physicochemical characteristics improves the number of truly covariant pairs. However, there are many reported aa properties and the selection of the right ones presents a difficult choice. Interestingly, Atchley et al. [46] used multivariate statistical analyses on 494 aa properties [47] to produce a small set of highly interpretable numeric patterns of aa variability that can be used in a wide variety of analyses directed toward understanding the evolutionary, structural, and functional aspects of protein variability. This transformation summarizes the high level of redundancy in the original physicochemical attributes and produces much smaller, statistically independent, and well conditioned variables for subsequent statistical analysis [48]. The resultant factors are linear functions of the original data, fewer in number than the original, and reflect clusters of covarying traits that describe the underlying structure of the variables [46].

Factor analysis of the highly intercorrelated aa attributes resulted in five factors, a reduction in dimensionality of two orders of magnitude from the original 494 properties [46]. POLARF1 reflects polarity and simultaneous covariation in portion of exposed residues versus buried residues, non-bonded energy versus free energy, number of hydrogen bond donors, polarity versus non-polarity, and hydrophobicity versus hydrophilicity. HELIXF2 is a secondary structure factor. There is an inverse relationship of relative propensity for various aa in various secondary structural configurations, such as a coil, a turn, or a bend versus the frequency in an α-helix. SIZEF3 relates to molecular size or volume with high factor coefficients for bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight. CODONF4 reflects relative aa composition in various proteins and the number of codons for each aa. These attributes vary inversely with refractivity and heat capacity. CHARGEF5 refers to electrostatic charge with high coefficients on isoelectric point and net charge. Atchley et al. [48] showed how the transformation into one of the five multidimensional factors of physicochemical properties was useful in the analysis of Basic Helix-Loop-Helix proteins that bind DNA.

2.4 Multi-response Permutation Procedure (MRPP)

MRPP is a non-parametric permutation test for testing the hypothesis of no difference between two or more groups of entities [49]. Permutation tests represent the ideal situations where one can derive the exact probabilities associated with a test statistic, rather than approximate values obtained from common probability distributions, such as t, F and X^2 [50]. In the majority of studies, the population distribution is unknown and assuming a normal distribution is inappropriate for many biological datasets, which often are skewed, discontinuous, and multi-modal. The distance-functions that form the basis of the MRPP are used to detect differences in distributions, sensitive to both dispersion (variation) and shifts in central tendency (median) [51]. MRPP was used to test differences between the physicochemical properties of the two groups of MHC-binding peptides, which was performed by the program BLOSSOM [51].

2.5 Discriminant analysis

Discriminant analysis is a statistical approach that defines the latent structure of between-groups covariation and determines the subset of attributes that best separate a set of a priori defined groups [52]. We used stepwise discriminant analysis (SWDA), as described by Atchley et al [48], to rank the 40 transformed variables of the MHC-binding peptides (5 physicochemical factors x 8 aa sites) in terms of their ability to discriminate between the binding and non-binding peptides. A step-up variable selection procedure begins with no variables in the model and then a variable is added that contributes most to discriminating power of the model, as measured by Wilks' lambda likelihood ratio criterion [52]. The procedure continues adding the next best discriminating variable until the p value of the F statistic was higher than 0.01. Then, we used Canonical variate analysis (CVA) to predict the membership of each peptide using the best discriminating variables of the SWDA model. A Leave-one-out cross validation procedure was employed, where each case is classified by the functions derived from all cases other than that case. All these procedures were performed with the program SPSS 15.0 [53].

2.6 Physicochemical correlation between aa sites

Bioinformatics methods for detecting correlated mutations consist of two main steps: (i) alignment of homologous sequences and (ii) identification of pairs of columns in the alignment in which there is a statistically significant tendency for mutations in one column to be accompanied by corresponding and usually different mutations in the other column [54]. In the present study, a modified version of a recent algorithm [55] was used to analyze pair-wise relationships between aa sites. The approach is based on estimation of the correlation coefficient between the values of a physicochemical parameter at a pair of positions of sequence alignment. When the correlation coefficient between two sites is negative, an increase in the value of a property at position i will make more likely a substitution at position j that will result in a decrease in the value of the property, which suggests a net value compensatory substitution. When the correlation coefficient is positive, it may be assumed that substitutions keep constant the difference between the property values of two residues. All statistical analyses, calculations and randomizations of this study were performed using MATLAB [56] unless stated otherwise.

If the sequences are effectively unrelated, then the pairs of positions with a significant covariation must have structural or functional links. Sequences are unrelated if the relationships by descent have been lost and there is no longer a significant phylogenetic signal or the sequences were obtained by in vitro selection (such as the dataset of MHC-binding peptides) [55]. There are three different sources of covariation in related biological sequences (such as the dataset of HCV sequences): (i) chance, (ii) common ancestry, and (iii) structural or functional constraints. Effectively discriminating among these underlying causes is a difficult task with many statistical and computational difficulties [57]. There are many methods to test whether a correlation value reflects a significant association (possibly due to structural and functional constraints), or results from evolutionary history and stochastic events (background covariation) [14] but no single method has demonstrated general utility or achieved widespread acceptance [32]. We used the following four criteria to define the pairs of significantly correlated pairs of sites.

(i) Each one of the two sites has an entropy higher than 0.2370, which is 10% of the highest entropy found in the HCV polyprotein. Only 448 aa sites of the 3010 aa sites of the HCV polyprotein are above this entropy cutoff. This cut-off was chosen because prior modeling of protein coevolution showed that it is difficult to identify sites which are coevolving if they are highly conserved [32, 33].

(ii) A permutation procedure was performed, whereby the aa at each site in the sequence alignment was vertically shuffled. Ten thousand random alignments were created this way, simulating the distribution of correlation values under the null hypothesis that substitutions of aa at two sites are statistically independent. For each physicochemical factor, a pair of sites was considered significantly correlated if its correlation value in the observed dataset was higher than the correlation value for those two sites in any of the random datasets (p = 0.0001). We addressed the multiple comparisons problem with the False Discovery Rate approach, which controls the expected proportion of false positive results [58]. The False Discovery Rate in our study has a q-value of 0.00035 for the dataset of MHC-binding peptides and 0.00506 for the dataset of viral sequences.

(iii) Related sequences (such as the dataset of HCV sequences) are part of a hierarchically structured phylogeny and, therefore, for statistical purposes, cannot be regarded as being drawn independently from the same distribution. We used the data weighting approach based on Felsenstein's method [59] in the calculation of the correlation values, which is based on the assumption that the lower the time of divergence of two sequences from their common ancestor, the higher is the covariation between these two sequences. The one-dimensional weights were calculated using a distance matrix among sequences built using the synonymous sites of the full HCV genome.

(iv) We used a modified version of the method of Martin et al [32] and Gloor et al [33]. This method makes the assumption that each position in a multiple sequence alignment is affected equally by background correlation, and that the majority of positions in the alignment covary only because of common ancestry. On the basis of these assumptions, each alignment is used as its own null model for the identification of covarying positions. A critical correlation threshold was calculated using the value of the Student's distribution at a given significance level (p = 0.001) with a sample size of 114 sequences, following Afonnikov et al [23].

3 Results

3.1 Dataset of MHC-binding peptides

There are two structurally distinct groups of peptide sequences in the dataset: those which bind to a MHC class I Kb molecule and those which do not. The dataset was transformed to the five physicochemical factors and it was found that these two groups occupy a significantly different region of the physicochemical space (p = 0.0001; MRPP). Now that we know that the two classes of peptides have significant physicochemical differences, it is important to understand the causes of these differences. We used SWDA to rank the 40 transformed variables of the MHC-binding peptides in terms of their ability to discriminate between binders and non-binders. The results indicate that some positions and factors contribute much more strongly than others in separating the data (Table 1), positions 5 and 8 being the most important, especially regarding the physicochemical properties summarized by POLARF1 and HELIXF2. This model includes 8 variables that account for 70.28% of the variability, and allow the correct classification of 91.3% of the peptides (90.3 of cross-validated cases are correctly classified) (Table 2).

Table 1. Discriminant analysis (SWDA) of the MHC-binding peptides

Step	Variable	Residual Variance	Significance of F
1	SIZEF3_P5	0.6310	2.5751E-06
2	HELIXF2_P8	0.4851	1.4468E-07
3	POLARF1_P8	0.4229	1.0112E-06
4	POLARF1_P5	0.3722	2.3665E-15
5	CODONF4_P5	0.3298	1.6459E-10
6	SIZEF2_P1	0.3117	2.2650E-04
7	SIZEF2_P3	0.3048	3.5888E-03
8	POLARF1_P2	0.2973	6.6078E-03

Table 2. Classification results of the CVA 8-variable model.

		Predicted binders	**Predicted non-binders**
	Observed binders	92.82%	7.18%
Original	**Observed non-binders**	10.85%	89.15%
	Observed non-binders	92.27%	7.73%
Cross-validated	**Observed non-binders**	12.40%	87.60%

What physicochemical characteristics are the sequences of binders keeping constant? Which pairs of sites are highly associated in order to conserve affinity? There are eight pairs of positions with significant correlations (links) in the set of binders (Table 3), but none in the set of non-binders (Fig. 2). Almost all positions in the MHC-binding dataset have one or more links, except position 8. However, position 8 has a link to position 7 with a lower significance (p = 0.0012). Position 3 has the highest number of links, followed by positions 1 and 5. The correlated positions in binding sequences keep constant factors POLARF1, SIZEF3 and CHARGEF5 of the peptide. These results suggest that our simple covariation analysis is useful for finding pairs of sites that are crucial to keeping the structural conformation of peptides.

Table 3. Significant physicochemical correlations at one or more factors (p = 0.0001) in binders.

i	j	POLARF1	HELIXF2	SIZEF3	CODONF4	CHARGEF5
1	3			-0.2823		
1	5	0.2701				
1	6			-0.3889		-0.3520
2	3	0.4408		0.3357		0.2821
2	5			-0.3087	-0.3617	-0.3019
3	5	0.3275	0.4030			
3	7	-0.2822				
4	6				0.3119	

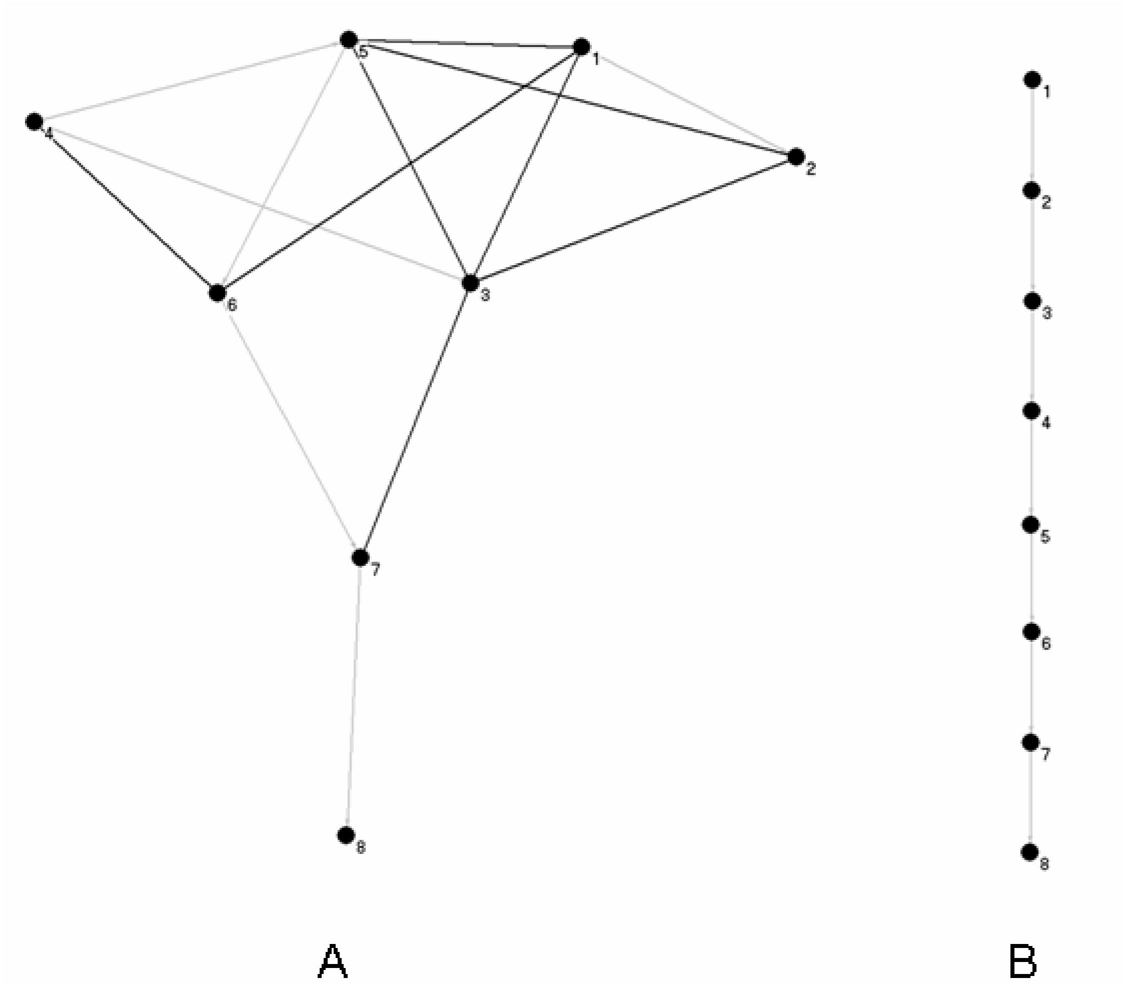


Fig. 2. Graph of the relationships between sites in the binders (A) and non-binders (B). Contiguous sites in the sequence are connected *(grey lines)* and sites with a significant physicochemical correlations at one or more factors (p = 0.0001) are also connected *(black lines)*. There are eight pairs of positions with significant correlations in the set of binders, but none in the set of non-binders.

3.2 HVR1

We found five links involving eight different sites in the middle of HVR1 (Table 4 and Fig. 3). The site with the highest number of links is 402 (3 links), suggesting an important role in keeping HVR1 structure and/or function. Changes in the HVR1 are correlated in a way that keeps constant the POLARF1 and SIZEF3 of the segment. The results suggest that there are three physicochemical traits or conditions that have been selected in the HVR1: the first in a cluster of sites (391, 396, 402 and 403) related to POLARF1 where site 402 seems critical; the second (390 and 393) related to SIZEF3, and the third (394 and 397) related to POLARF1.

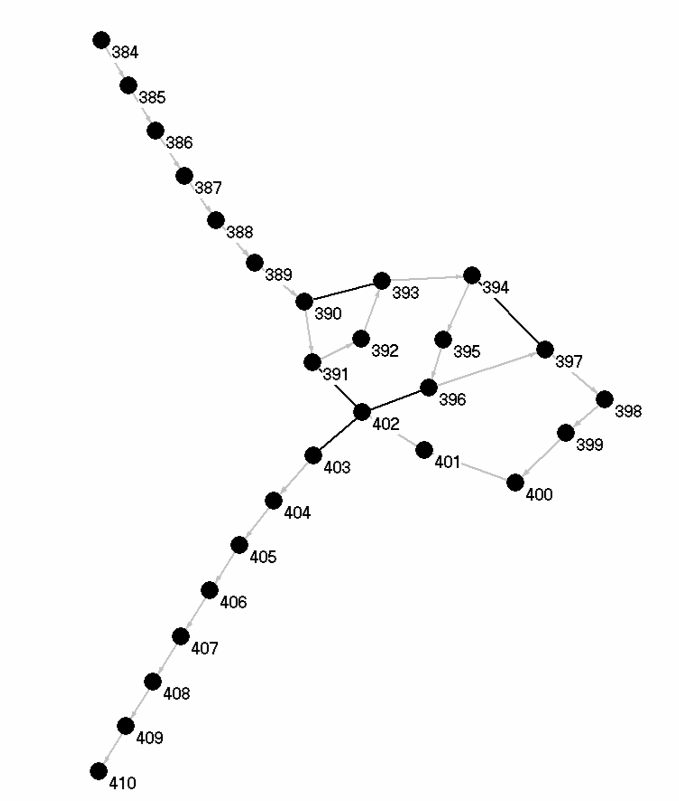


Fig. 3. Graph illustrating relationships between sites in the HVR1. Contiguous sites in the sequence are connected *(grey lines)* and sites with a significant physicochemical correlation at one or more factors (p = 0.0001) are also connected *(black lines)*.

Table 4. Significant physicochemical correlations at one or more factors (p = 0.0001) in HVR1.

i	j	POLARF1	HELIXF2	SIZEF3	CODONF4	CHARGEF5
390	393			-0.3063		
391	402	0.4090				
394	397	-0.3759				
396	402	0.3522				
402	403				-0.5505	

4 Discussion

Knowledge about the determinants of binding to MHC molecules is very useful for the development of predictive tools that help choose peptides for employment in immunological therapies or inclusion in vaccines intended to elicit T-cell cytotoxic activity. It has been shown that some aa residues occur at a high frequency at specific positions in the peptide, termed anchor positions [36, 37]. For the MHC class I molecule (K^b) the anchor positions are 3, 5 and 8, with preferences for aa tyrosine or phenylalanine in positions 3 or 5 and a hydrophobic aa in position 8. However, binding is known to be influenced by both the presence of secondary anchor positions and interactions between aa within the peptide [37]. Interestingly, we found that positions 3, 5 and 8 are the best positions in discriminating between binders and non-binders, in agreement with their role as anchor positions. We also found that there is a high level of covariation between all positions in MHC binding peptides but none in the set of non-binding peptides. This observation clearly suggests that the ability to bind to MHC creates constraints on the sequence variability of the binding peptides, where changes at some positions are coordinated with changes at other positions in order to maintain binding capacity. The covariation level of position 8 is very low, suggesting that its contribution to MHC binding is more independent of the other sites, even though it is an anchor position and is the most important position discriminating between binders and non-binders. The aa distribution of position 8 is very different in the binding and non-binding dataset, with significant differences in the average POLARF1 ($p = 0.0001$; MRPP) and HELIXF2 ($p = 0.0001$; MRPP) of the two groups. These results suggest that there are two physicochemical traits or conditions affecting binding in these 8-mer peptides: the first is related to positions 1-7 (conserving POLARF1, SIZEF3 and CHARGEF5) and the second is related with position 8 (POLARF1 and HELIXF2), which is less dependent of the other positions but very important to define the ability to bind.

We also studied the sequence variability of HVR1 in order to establish if there is a mechanism underlying the selection of this subset of aa sequences. The results suggest that the HVR1 segment has to keep a specific structure and/or function and that natural selection left a mark in the sequence variability in the form of coordinated substitutions. The high number of coordinated substitutions and their contribution to the maintenance of some physicochemical values provide additional proof to the conservation of conformational motifs in the HVR1, for which there is previous experimental evidence [43]. This conservation is consistent with strong selective constraints previously found on HVR1 heterogeneity [42, 60, 61], suggesting that this segment has an important function in virus replication, rather than merely being a variable region of the genome that acts as an antigenic decoy [42, 60]. The results suggest that the physicochemical properties POLARF1 and SIZEF3 have been selected in the HVR1. The high number of coordinated substitutions and their contribution to the integrity of these physicochemical properties provide an additional proof to conservation of conformational motifs in the HVR1. Covariation analyses can be important in identifying sites that may change the phenotype of a protein, and they could be used as a tentative map for researchers to define functional domains in the protein through mutational analysis. For instance, covariant sites could be used as a guide for rational selection of sets of sequences for inclusion in a mixture of peptides for vaccine design. Therefore, by selecting sequences which include pairs of aa that are highly predictive of each other, important classes of sequences that are structurally or functionally related may be identified. Thus inclusion of peptides with highly covariant aa may be a useful strategy for designing broadly-reactive vaccines [28].

The detection of coordinated substitutions among separate aa sites is fundamental to understanding protein structure and evolution. The process of selection (whether natural or *in vitro*) creates profound constraints on the sequence variability, keeping constant the structure or function of the protein. We found the consequences of these constraints in the form of covariant pairs of sites and the conservation of specific physicochemical properties.

Acknowledgments. The authors are grateful to Dr. Chong-Gee Teo (Centers for Disease Control and Prevention, Atlanta, GA) and Dr. Robert Mitchell (La Trobe University, Melbourne, Australia) for valuable comments at different stages of the project.

References

1. Pollock, D. and W. Taylor, Effectiveness of correlation analysis in identifying protein residues. Protein Eng, 1997. 10(6): p. 647-657.
2. Chothia, C. and A. Lesk, Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. J Mol Biol, 1982. 160(2): p. 309-23.
3. Lesk, A. and C. C., Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. J Mol Biol, 1982. 160(2): p. 325-42.
4. Oosawa, K. and M. Simon, Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in Escherichia coli. Proc Natl Acad Sci USA, 1986. 83(18): p. 6930-4.
5. Altschuh, D., et al., Coordinated amino acid changes in homologous protein families. Protein Eng, 1988. 2(3): p. 193-9.
6. Bordo, D. and P. Argos, Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. . J Mol Biol. , 1990. 211(4): p. 975-88.
7. Mateu, M. and A. Fersht, Mutually compensatory mutations during evolution of the tetramerization domain of tumor supressor p53 lead to impaired hetero-oligomerization. Proc Natl Acad Sci USA, 1999. 96: p. 3595–3599.
8. Lim, W. and R. Sauer, Alternative packing arrangements in the hydrophobic core of lambda repressor. Nature, 1989. 339(6219): p. 31-6.
9. Lim, W., D. Farruggio, and R. Sauer, Structural and energetic consequences of disruptive mutations in a protein core. Biochemistry, 1992. 31(17): p. 4324-33.
10. Baldwin, E., et al., The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. Science, 1993. 262(5140): p. 1715-8.
11. Govindarajan, S., et al., Systematic variation of Amino acid substitutions for stringent assesment of pairwise covariation. J Mol Biol, 2003. 328: p. 1061-1069.
12. Clarke, N., Covariation of residues in the homeodomain sequence family. Protein Sci, 1995. 4(11): p. 2269-78.
13. Voigt, C., et al., Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci USA, 2001. 98: p. 3778–3783.
14. Atchley, W., et al., Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol, 2000. 17(1): p. 164-78.
15. Fukami-Kobayashi, K., D. Schreiber, and S. Benner, Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol, 2002. 319: p. 729–743.
16. Göbel, U., et al., Correlated mutations and residue contacts in proteins. Proteins, 1994. 18(4): p. 309-17.
17. Neher, E., How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA, 1994. 91(1): p. 98-102.
18. Shindyalov, I., N. kolchanov, and C. Sander, Can three dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng, 1994. 7: p. 349-358.
19. Taylor, W. and K. Hatrick, Compensating changes in protein multiple sequence alignments. Protein Eng, 1994. 7(3): p. 341-8.
20. Benner, S., et al., Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Rev., 1997. 97: p. 2725-2844.
21. Nagl, S., J. Freeman, and T. Smith, Evolutionary constraint networks in ligand-binding domains: an information-theoretic approach. Pac Symp Biocomput, 1999: p. 90-101.
22. Larson, S., A. Di Nardo, and A. Davidson, Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. J Mol Biol, 2000. 303(3): p. 433-46.
23. Afonnikov, D., D. Oshchepkov, and N. Kolchanov, Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics, 2001. 17(11): p. 1035-46.
24. Nemoto, W., et al., Detection of pairwise residue proximity by covariation analysis for 3D-structure prediction of G-protein-coupled receptors. Protein J, 2004. 23(6): p. 427-35.
25. Wang, L., Covariation analysis of local amino acid sequences in recurrent protein local structures. J Bioinform Comput Biol, 2005. 3(6): p. 1391-409.
26. Shackelford, G. and K. Karplus, Contact prediction using mutual information and neural nets. Proteins, 2007. 69(Suppl 8): p. 159-64. .
27. Altschuh, D., et al., Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol, 1987. 193(4): p. 693-707.
28. Korber, B., et al., Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA, 1993. 90(15): p. 7176-80.
29. Gilbert, P., V. Novitsky, and M. Essex, Covariability of selected amino acid positions for HIV type 1 subtypes C and B. . AIDS Res Hum Retroviruses, 2005. 21(12): p. 1016-30.

30.Kolli, M., S. Lastere, and C. Schiffer, Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate. Virology, 2006. 347(2): p. 405-9.

31.Chelvanayagam, G., et al., An analysis of simultaneous variation in protein structures. Protein Eng, 1997. 10(4): p. 307-316.

32.Martin, L., et al., Using information theory to search for co-evolving residues in proteins. Bioinformatics, 2005. 21(22): p. 4116-24.

33.Gloor, G., et al., Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry, 2005. 44(19): p. 156-65.

34.Poon, A. and L. Chao, The rate of compensatory mutation in the DNA bacteriophage phiX174. Genetics, 2005. 170(3): p. 989-99.

35.Yeang, C. and D. Haussler, Detecting coevolution in and among protein domains. PLoS Comput Biol. , 2007 3(11): p. e211.

36.Milik, M.S., D. Brunmark, A. Yuan, L. Vitiello, A. Jackson, M. Peterson, P. Skolnick, J. Glass, C., Application of an artificial neural network to predict specific class I MHC binding peptide sequences. Nat Biotechnol, 1998. 16(8): p. 753-6.

37.Segal, M., M. Cummings, and A. Hubbard, Relating amino acid sequence to phenotype: analysis of peptide-binding data. Biometrics, 2001. 57(2): p. 632-42.

38.Alter, M., Epidemiology of hepatitis C virus infection. World J Gastroenterol, 2007. 13(17): p. 2436-41.

39.Alberti, A., L. Chemello, and L. Benvegnu, Natural History Of Hepatitis C. J Hepatol, 1999. 31(Supp1): p. 17–24.

40.Bowen, D. and C. Walker, Adaptive immune responses in acute and chronic hepatitis C virus infection. Nature, 2005. 436: p. 946-952.

41.Choo, Q., et al., Isolation Of A Cdna Clone Derived From A Bloodborne Non-A, Non-B Viral Hepatitis Genome. Science, 1989. 244: p. 359–62.

42.Smith, D., Evolution of the hypervariable region of hepatitis C virus. J Viral Hepat, 1999. 6(Suppl1): p. 41–46.

43.Mondelli, M., et al., Hypervariable region 1 of hepatitis C virus: immunological decoy or biologically relevant domain? Antiviral Res, 2001. 52(2): p. 153-9.

44.Kuiken, C., et al., The Los Alamos hepatitis C sequence database. Bioinformatics, 2005. 21(3): p. 379-84.

45.Thompson, J., D. Higgins, and T. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994. 22(22): p. 4673-80.

46.Atchley, W., et al., Solving the protein sequence metric problem. Proc Natl Acad Sci USA, 2005. 102(18): p. 6395-400.

47.Kawashima, S. and M. Kanehisa, AAindex: amino acid index database. Nucleic Acids Res, 2000. 28: p. 374.

48.Atchley, W. and J. Zhao, Molecular architecture of the DNA-binding region and its relationship to classification of basic helix-loop-helix proteins. Mol Biol Evol, 2007. 24(1): p. 192-202.

49.McCune, B. and J. Grace, Analysis of ecological communities. 2002, Gleneden Beach: MjM Software Design.

50.Cai, L., Multi-response Permutation Procedure as An Alternative to the Analysis of Variance: An SPSS Implementation. Department of Psychology, University of North Carolina, 2004.

51.Cade, B. and J. Richards, User Manual For BLOSSOM Statistical Software. Midcontinent Ecological Science Center US Geological Survey Fort Collins, Colorado, 2001.

52.Johnson, R. and D. Wichern, Applied multivariate statistical analysis. 2002, Upper Saddle River, NJ: Prentice Hall.

53.SPSS 15.0 for windows. 2006, SPSS Inc: Chicago IL.

54.Noivirt, O., M. Eisenstein, and A. Horovitz, Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng, 2005. 18(5): p. 247-253.

55.Afonnikov, D. and N. Kolchanov, CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. Nucleic Acids Res, 2004. 32: p. W64-W68.

56.MathWorks, T., MATLAB. 2007: Natick, MA.

57.Wollenberg, K. and W. Atchley, Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci USA, 2000. 97(7): p. 3288-3291.

58.Benjamini, Y. and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical Society, Series B, 1995. 57(1): p. 289-300.

59.Felsenstein, J., Phylogenies and the comparative method. Am Nat, 1985. 125: p. 1–15.

60.McAllister, J., et al., Long-term evolution of the hypervariable region of hepatitis C virus in a common-source-infected cohort. J Virol, 1998. 72(6): p. 4893-905.

61.Sheridan, I., et al., High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. J Virol, 2004. 78(7): p. 3447-54.