CD8+ T-cell Epitope Presentation on MHC-1 as a Probability Space

Rohin E. McIntosh

PACM Independent Work
Department of Physics
Princeton University

Advised by Professor Curtis Callan and Dr. Barbara Bravi
Second Reader Prof. Jason Fleischer

This paper represents my own work in accordance with University regulations.

/s/ Rohin E. McIntosh

Abstract

With continually growing access to immune epitope sequence data, machine learning approaches promise to greatly expand our theoretical understanding of the immune system. Following a Restricted Boltzmann Machine peptide presentation prediction algorithm (RBM-MHC), we propose an information theoretic approach to the characterization of CD8+ T-cell epitope presentation by MHC-1 molecules. For a given HLA-1 allele, RBM-MHC learns a probability distribution for the corresponding epitope presentation space. Given this form, we demonstrate that tools from information theory and thermodynamics, in particular Shannon entropy and the Jensen-Shannon Divergence, provide important and novel insights into HLA allele diversity and overlap as well as expand our ability to quantitatively characterize and compare HLA alleles.

Contents

1 Introduction

The increasing availability of immunopeptidomic datasets produces constant novelty in our understanding of peptide repertoires binding to given Major Histocompatibility Complexes (MHCs). An outstanding open question is how to describe, through a theoretical framework, the diversity within the same repertoire and across repertoires, given also the huge HLA polymorphism. This poses the need for approaches that extrapolate from available data, e.g. model-based approaches, to give a quantitative characterization of repertoire size, complexity and overlap. Attempts undertaken so far rely on counting peptides in databases to estimate repertoire sizes [1, 2] and on the correlation between sequence motifs to quantify overlap between repertoires [2]. Here we propose alternative tools based on inferring probability distributions - and in particular probability distributions that account for correlations between sequence positions.

Using epitope sequence data from the Immune Epitope Database and Analysis Resource (IEDB) [3], a Restricted Boltzmann Machine (RBM) is used to learn a probability distribution from peptides presented by a given HLA-1 allele. Analyzing the diagnostic ability of these models, we have previously shown that the RBM is a more powerful classifier than other existing algorithms [4]. In addition to outperforming other classifiers in its predictive ability, the model is also capable of generating samples from the learned distribution.

For each HLA-1 allele, the RBM model learns a probability distribution for the space of presented epitopes in the form of the Boltzmann distribution. Taking inspiration from statistical mechanics and information theory, we currently seek to investigate the properties and applications of corresponding thermodynamic quantities. In particular, for any given HLA-1 allele, we can calculate a Shannon entropy for the presentation space. From an information theoretic perspective, the entropy describes the information we have about which states are likely to be sampled. We then hypothesize that the entropy informs us about the diversity of an allele's epitope presentation space. By comparing our results to those obtained from a naively generated probability distribution where only frequencies of single amino acid residues are considered (the independent model), we can quantify the information stored by the couplings between residues as the difference between the two entropies, called the multiinformation [5].

An individual's HLA type has a medically important role in the determination of transplant compatibility. Because of the huge variety of HLA genotypes in the human population, perfect donor-recipient matches for transplants are often impossible to find [6]. Therefore, it is vital that we have an accurate theoretical understanding of which alleles are most similar to one another. The form of the RBM model as a probability distribution suggests a simple method for quantitatively comparing an arbitrary number of alleles or allele sets in the form of the Jensen-Shannon divergence (JSD). Using this quantity, we investigate both the diversity of alleles within the same HLA type (HLA-A, HLA-B, and HLA-C) as well as the pairwise similarities between alleles. Throughout this paper, our results are obtained from RBM models trained on the 98 HLA-1 alleles for which the most data is available in the IEDB database.

2 Restricted Boltzmann Machines

In order to characterize an HLA-1 allele's epitope presentation space, we utilize a Restricted Boltzmann Machine (RBM): an energy-based machine learning model capable of learning a probability distribution from sample data. RBMs provide an elegant and novel machine learning approach to predict peptide antigen presentability that is critically important to immune responses. In previous work, we demonstrated an RBM algorithm capable of learning probability distributions of amino acids [7] and applied this algorithm to the classification of HLA-1 alleles [4]. In both its predictive and generative power, this algorithm has proven superior to other existing models.

A RBM is a two layer neural network with a visible and a hidden layer. In our implementation, the visible layer takes a peptide of fixed length (9-mer for HLA-1) and the hidden layer represents features of the training data learned by the RBM. The model itself is a function that computes an energy for any possible

9-mer peptide from the equation,

$$E(v,h) = -\sum_{i=1}^{N} g_i(v_i) + \sum_{j=1}^{M} \mathcal{U}_j(h_j) - \sum_{i=1}^{N}\sum_{j=1}^{M} h_j W_{i,j} v_i \tag{1}$$

where N is the number of visible nodes, M is the number of hidden nodes, v_i are the visible nodes, h_j are the hidden nodes, $W_{i,j}$ are the weights, and $g_i(v_i)$ and $\mathcal{U}_j(h_j)$ represent the biasing potentials of the visible and hidden units, respectively [8]. A probability distribution is then constructed in the form of the Boltzmann distribution,

$$p(v) = \sum_{j=1}^{M} \frac{e^{-E(v,h_j)}}{Z}. \tag{2}$$

where the hidden nodes have been summed over to describe the probability that the machine samples a given input vector v and Z is the partition function defined as the sum over all possible states. Training the RBM then amounts to maximizing this probability (minimizing the energy) for the peptides in provided in the training set. The details of the algorithm and training are described in [8].

3 Entropy

3.1 Definition

For each HLA-1 allele, the RBM model learns a probability distribution for the space of presented epitopes in the form of the Boltzmann distribution. Taking inspiration from statistical mechanics and information theory, we investigate the properties and applications of corresponding thermodynamic quantities. The most important of these quantities is the entropy which is defined as the expectation value of the negative log-probability,

$$S_{RBM} = -\sum_{x\in\Omega} P(x) \ln P(x) \tag{3}$$

for probability distribution P defined over the sample space Ω. From an information theoretic perspective, the entropy describes the information we have about which states are likely to be sampled. For our application, this means that a MHC molecule which presents all possible peptides equally well, will have the maximum entropy ($S_{max} = \ln \Omega$). Conversely, one that can only present one peptide will have zero entropy.

3.2 Independent Model

A convenient model for comparison is the independent model. For a given HLA-1 allele's presentation space, we define the probability of a sequence in this model as the product of the amino acid frequencies calculated from the data at each individual residue. This is similar to the probability distribution learned by the RBM except that it ignores all features corresponding to coupling between residues. As a result, the independent entropy is always larger than the RBM entropy. Since the residues are independent, the entropy of this model can be computed exactly as,

$$S_{ind} = -\sum_{i=1}^{9} \sum_{j=1}^{20} P_i(x_j) \ln P_i(x_j) \tag{4}$$

where $P_i(x_j)$ is the frequency of amino acid x_j at residue i. The difference between the independent entropy and the RBM entropy represents the information learned from the couplings between residues and is called the multiinformation [5].

3.3 Individual Alleles

For each of the 98 HLA-1 alleles with trained RBM models, we calculated the entropies for both the RBM models and the independent models. Our results are shown in figure 1. The error bars for both entropies represent the standard deviation in the negative log probability (or the standard deviation in the self-information). While we suspect that correlations between allele entropy and general CD8+ T-cell activity can be made, most studies to date correlate single alleles with specific diseases providing little information about total immune activity for comparison. For each allele, we also present the multiinformation in figure 2 illustrating an improvement of at least 0.5 nats over the independent model for most alleles.

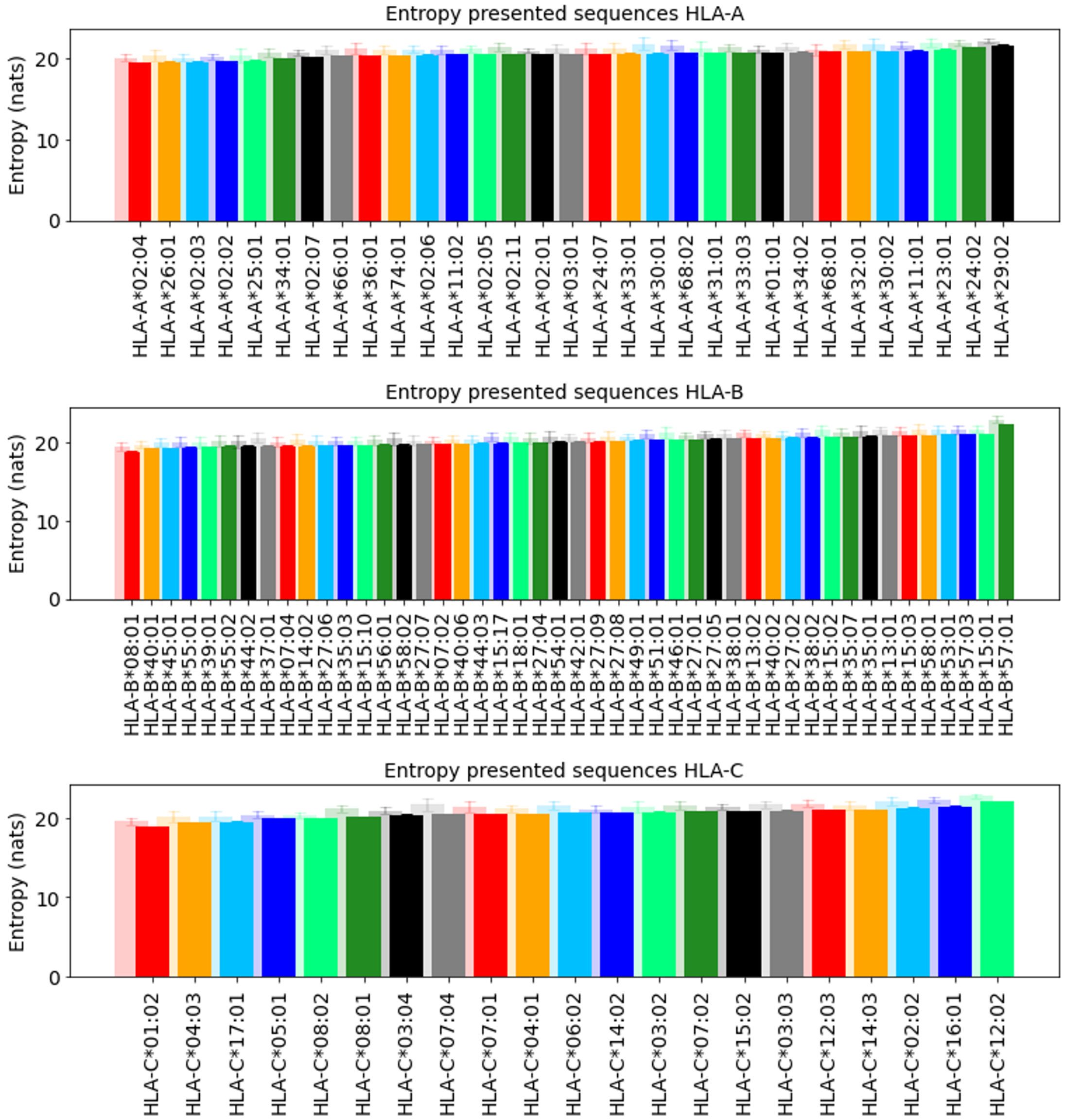

Figure 1: RBM (dark) and independent model (light) entropy for each of the 98 HLA-1 alleles. Error bars represent the standard deviation in the self-information.

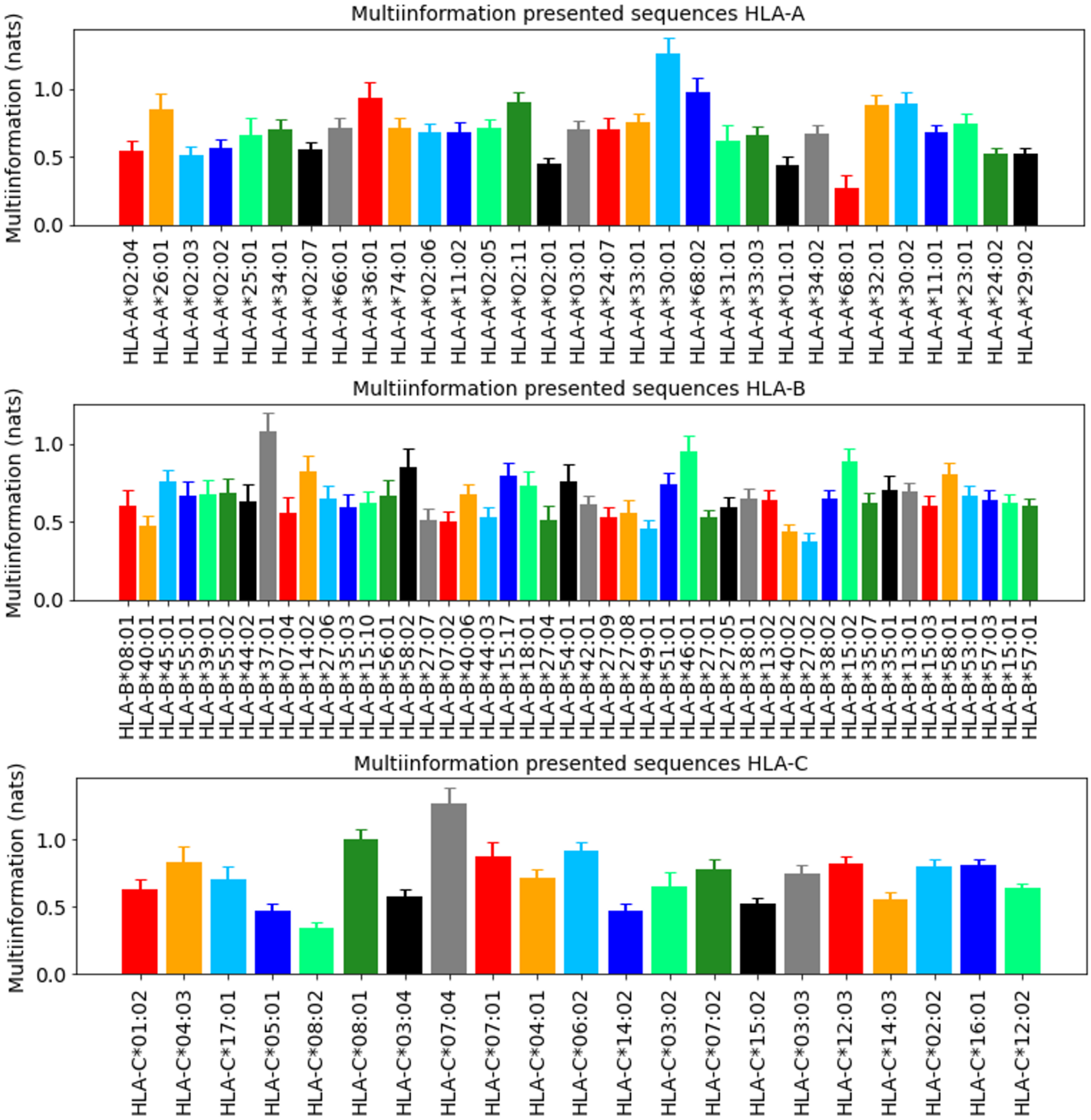

Figure 2: Multiinformation for each of the 98 HLA-1 alleles. Error bars are calculated from propagation of uncertainty.

4 Jensen-Shannon Divergence

4.1 Definition

Since for each allele we have inferred a probability distribution, it is useful to find some metric by which to compare these distributions. A common method for comparing probability distributions is the Kullback-Leibler divergence defined as,

$$D_{KL}(P||Q) \coloneqq \sum_{x\in\Omega} P(x)\ln\frac{P(x)}{Q(x)} \tag{5}$$

for probability distributions P and Q defined over the same space Ω. In order to symmetrize this quantity and ensure boundedness, we consider the Jensen-Shannon Divergence (JSD),

$$\begin{aligned} JSD(P||Q) \coloneqq& \frac{1}{2}D(P||\frac{(P+Q)}{2}) + \frac{1}{2}D(Q||\frac{(P+Q)}{2}) && (6) \\ =& \sum_{x\in\Omega} P(x)\ln\frac{2P(x)}{P(x)+Q(x)} + \sum_{x\in\Omega} Q(x)\ln\frac{2Q(x)}{P(x)+Q(x)}. && (7) \end{aligned}$$

This quantity can also be defined as the mutual information between a binary indicator variable and the mixture distribution between P and Q. In this case, the binary indicator variable returns either P or Q with equal probability, although this may be made more general, where for instance, P and Q are returned with different weights. This definition suggests the information theoretic interpretation that the Jensen-Shannon Divergence is the average information that a sample gives about the distribution from which it came. In fact, the square root of the JSD is a metric, since it satisfies the triangle inequality, and is termed the Jensen-Shannon Distance (JSDist). Therefore, the JSD functions as a squared distance between probability distributions [9].

4.2 JSD in Epitope Presentation Spaces

Applied to the epitope presentation spaces of two HLA-1 alleles, the JSD provides useful information about how much these spaces overlap. If the JSD is maximal ($JSD_{max} = \ln 2$), this implies that for every sequence presentable by one MHC complex, there is 0 likelihood of it being presentable by the other complex. If the JSD is minimal ($JSD_{min} = 0$), we have that the two presentation spaces are identical. A higher JSD indicates a greater likelihood of cross-priming events: when epitopes are capable of being presented by both MHC complexes.

4.3 Interactions Learned by the RBM

The JSD can be generalized to compare any number of probability distributions $(P_1, P_2, ..., P_N)$ defined over the same space:

$$JSD(P_1, P_2, ..., P_N) := \frac{1}{N} \sum_{n=1}^{N} D(P_n || \frac{\sum_{n=1}^{N} P_n}{N}) \tag{8}$$

$$= \sum_{n=1}^{N} \sum_{x \in \Omega} P_n(x) \ln \frac{N P_n(x)}{\sum_{n=1}^{N} P_n}. \tag{9}$$

In figure 3 we computed this quantity for all combinations of alleles considering 10 different alleles of each of the three HLA-1 loci. Universally, we find that the JSD between RBM models is greater that that between their independent model counterparts. Since the independent model considers each amino acid residue independently, this difference represents the information contained by the couplings between amino acid residues that are learnt by the RBM model. Comparing between the three HLA-1 types, we can see that on average, HLA-B alleles have the greatest diversity amongst themselves followed by HLA-A and then HLA-C.

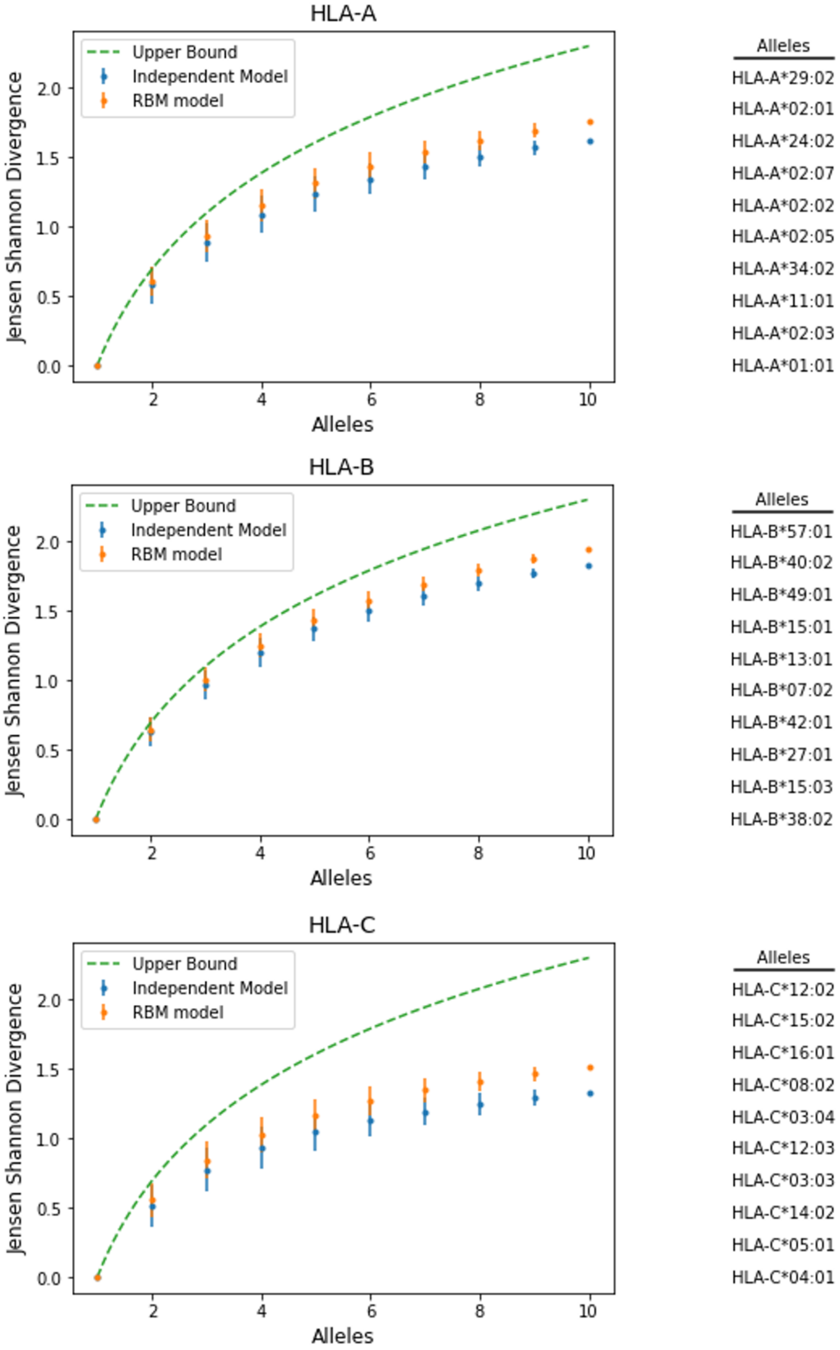

Figure 3: Average generalized Jensen-Shannon Divergence as a function of number of alleles considered. For each x-value, the Jensen-Shannon Divergence of all combinations of x alleles taken from the lists on the right are averaged over for the both the RBM and independent models. Error bars represent 1 standard deviation about the mean. JSD upper bound represents value at which all distributions are disjoint.

4.4 Pairwise Distance Between Alleles

To investigate the pairwise JSD landscape across all 98 HLA alleles, we computed the JSD for all pairs of alleles. In figure 4, we present the data grouped by HLA type. This clearly illustrates differences in diversity among alleles of the same type: HLA-C is the least diverse HLA type forming a prominent tight cluster while HLA-B is the most diverse with some alleles sharing very little in common with others in the same group. This supports our conclusions from the previous section investigating mutiallelic JSD. We also observe that in general, HLA-A and HLA-B alleles individually share more in common with HLA-C alleles than they do with each other. Individual group diversity suggests the previously proposed hypothesis that HLA-B evolutionary precedes the other two groups and that HLA-C is the youngest HLA-1 type [1]. However, the relative divergence between HLA-1 types A and B suggests, on the contrary, that HLA-A and HLA-B are both evolutionary derived from HLA-C. Additional methods used to pairwise compare individual alleles, such as the Grantham distance that quantitatively compares amino acid sequences by their physicochemical properties, produce similar qualitative results [1].

HLA alleles are also sub-classified into supertypes. To investigate the validity of these classifications and to test our own model, we also present a clustered JSD heat map (figure 5) where alleles are color coded by their supertype classifications as defined by [10, 11]. Our results demonstrate that, in general, alleles within the same supertype tend to cluster together. These results also suggest supertype placements for unclassified alleles, such as for HLA-A*34:01 and supertype A03. Accurate classification of similar alleles is important especially when the diversity of HLA alleles in the human population makes finding suitable transplant donors difficult. In the absence of a perfect match, usually only achieved when the donor is a relative, the donor and recipient's respective HLA types should be as similar as possible. In fact, a recent study demonstrated that HLA-B is one of the most important alleles to consider when a mismatch is inevitable [6]. This aligns with the result that HLA-B is the most diverse of the HLA-1 types.

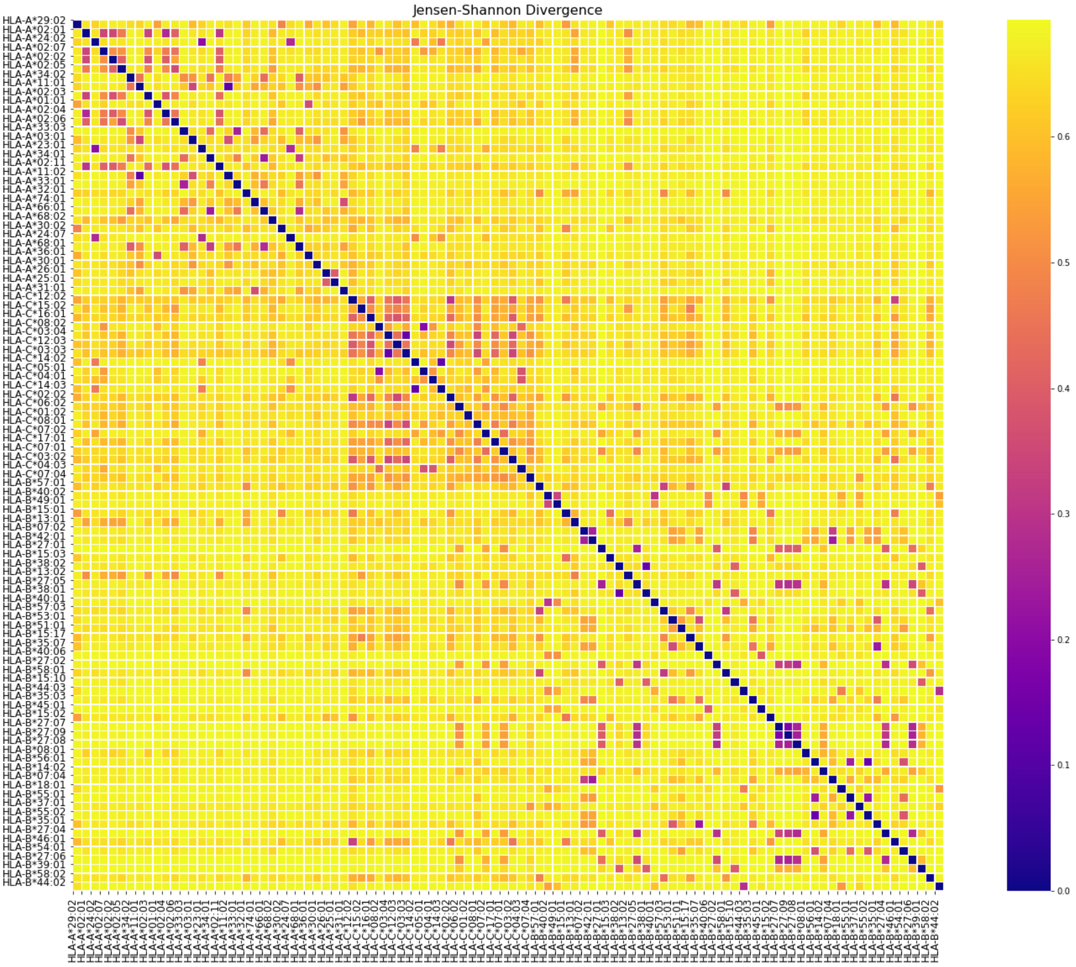

Figure 4: Jensen-Shannon divergence for all pairs of alleles grouped by HLA type (HLA-A, HLA-B, and HLA-C). Clustering provides clues as to the evolutionary origins of each HLA type.

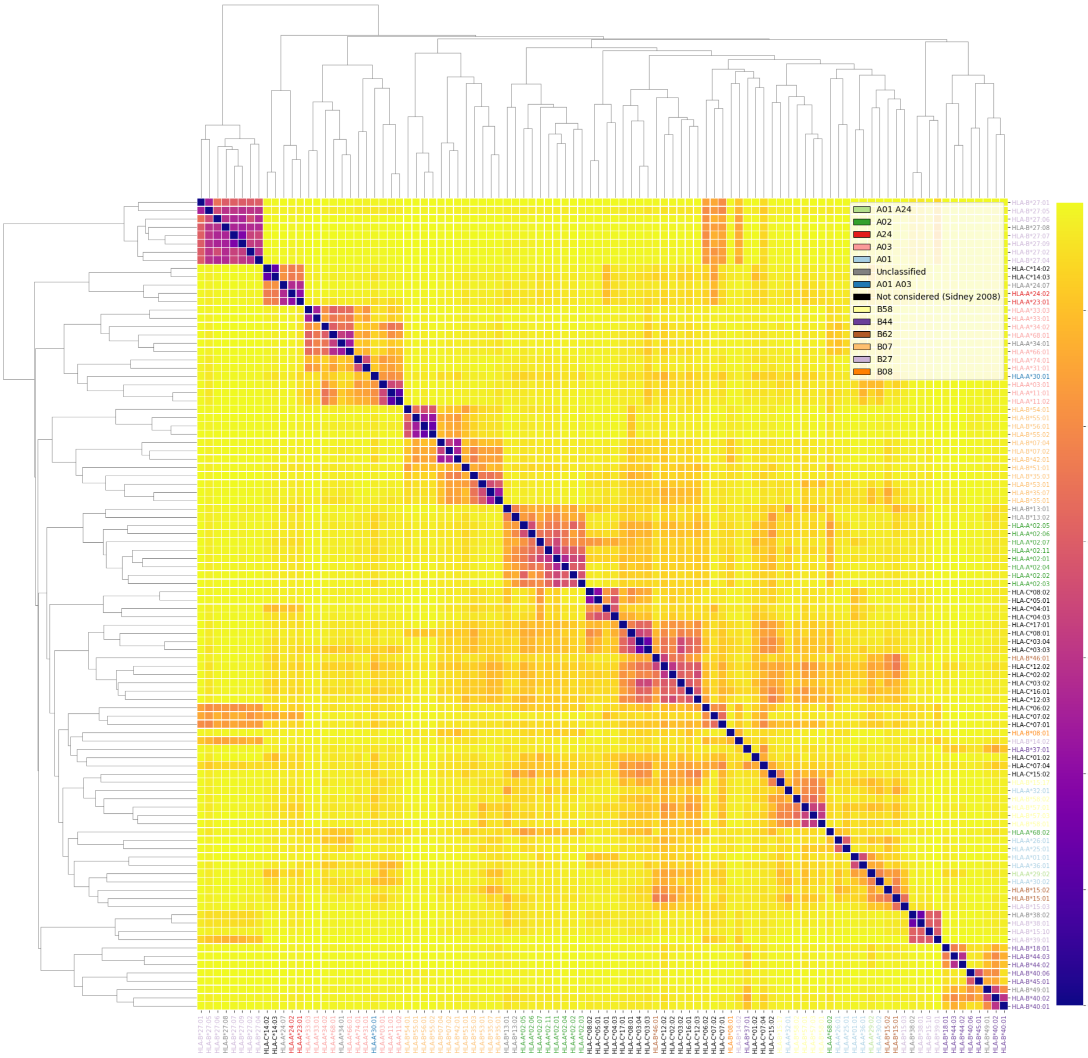

Figure 5: Jensen-Shannon divergence for all pairs of alleles. Heat map is clustered to group alleles with smaller distances. Alleles are color coded by their classification as defined by [10, 11].

5 Conclusion

Probability spaces for T-cell epitope presentation by MHC-1 molecules make an excellent model for the theoretical investigation of HLA-1 alleles. We have demonstrated that by using a Restricted Boltzmann Machine to learn these distributions from large data sets, we can quantitatively characterize properties both of individual alleles and similarities/differences between alleles and sets of alleles. The diversity of a presentation space is characterized by the RBM model entropy which may help inform the range of protection granted by an individual allele. We are also able to quantify the information stored in the couplings between amino acid residues using the multiinformation. In addition, the JSD provides us with a convenient method for comparing these probability distributions both pairwise and as sets of alleles. These comparisons provide important information about the diversity of alleles within sets (ex. HLA-A, supertypes, haplotypes) in addition to defining a metric with which to compare individual alleles. Quantification of pairwise comparisons in particular has important implications for characterizing donor/recipient transplant compatibility.

6 Acknowledgements

I thank my advisors, Professor Curtis Callan and Dr. Barbara Bravi for their wonderful guidance and for introducing me to the world of restricted Boltzmann machines. I also again thank Dr. Barbara Bravi for sharing her code and analysis and well as Dr. Jérôme Tubiana for his implementation of the RBM algorithm. The RBM code may be found on GitHub through the following link: https://github.com/jertubiana/ProteinMotifRBM. I am grateful to Professor Jason Fleischer for being my second reader. Lastly, I thank my parents, Professors Sumita Bhaduri-McIntosh and Michael McIntosh, and my sister, Uma McIntosh, for their encouragement and guidance.

References

[1] D. Chowell *et al.*, *Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy*, Nature Medicine **25** (2019).

[2] S. Sarkizova *et al.*, *A large peptidome dataset improves HLA class I epitope prediction across most of the human population*, Nature Biotechnology **38** (2020).

[3] R. Vita *et al.*, *The immune epitope database (IEDB) 3.0*, Nucleic Acids Research (2015).

[4] B. Bravi *et al.*, *Flexible machine learning prediction of antigen presentation for rare and common HLA-I alleles*, bioRxiv (2020).

[5] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, *Maximum entropy models for antibody diversity*, Proceedings of the National Academy of Sciences of the United States of America **107** (2010).

[6] E. W. Petersdorf *et al.*, *HLA-B leader and survivorship after HLA-mismatched unrelated donor transplantation*, Blood **136** (2020).

[7] J. Tubiana, S. Cocco, and R. Monasson, *Learning protein constitutive motifs from sequence data*, eLife (2019), 1803.08718.

[8] J. Tubiana and R. Monasson, *Emergence of Compositional Representations in Restricted Boltzmann Machines*, Physical Review Letters (2017), 1611.06759.

[9] D. M. Endres and J. E. Schindelin, A new metric for probability distributions, 2003.

[10] J. Sidney *et al.*, *Quantitative peptide binding motifs for 19 human and mouse MHC class i molecules derived using positional scanning combinatorial peptide libraries*, Immunome Research (2008).

[11] M. Di Marco *et al.*, *Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction Matrices*, The Journal of Immunology **199** (2017).