Accepted Manuscript

Title: Scrutinizing Human MHC Polymorphism: Supertype Analysis using Poisson-Boltzmann Electrostatics and Clustering

Authors: Shahzad Mumtaz, Ian T. Nabney, Darren R. Flower

PII: S1093-3263(17)30511-9
DOI: http://dx.doi.org/doi:10.1016/j.jmgm.2017.07.033
Reference: JMG 7000

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 30-6-2017
Accepted date: 25-7-2017

Please cite this article as: Shahzad Mumtaz, Ian T.Nabney, Darren R.Flower, Scrutinizing Human MHC Polymorphism: Supertype Analysis using Poisson-Boltzmann Electrostatics and Clustering, Journal of Molecular Graphics and Modellinghttp://dx.doi.org/10.1016/j.jmgm.2017.07.033

Scrutinizing Human MHC Polymorphism: Supertype Analysis using Poisson-Boltzmann Electrostatics and Clustering

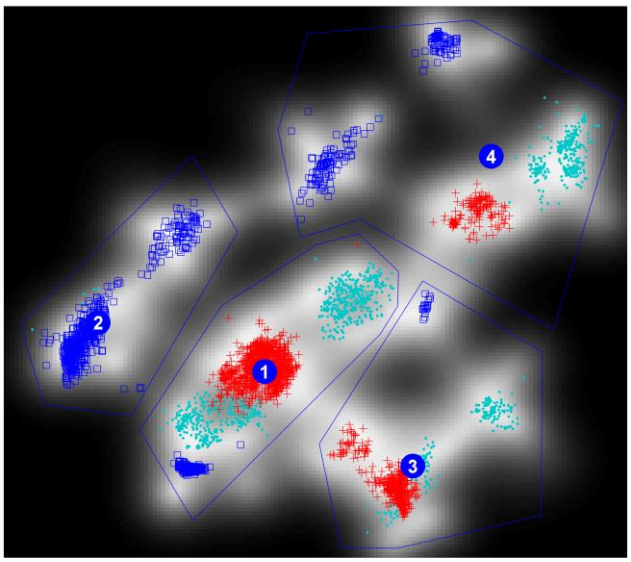
Shahzad Mumtaz[1], Ian T Nabney[1],

& Darren R Flower[2,*]

[1]School of Engineering and Applied Science, Aston University, Aston Triangle, Birmingham, United Kingdom, B4 7ET.

[2]School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham, United Kingdom, B4 7ET.

Graphical abstract



Highlights

- We outline new and improved approach to determining MHC supertype.
- This methodology combines Poisson Boltzmann electrostatics with state-of-the-art data reduction.

- The clustering we observe reconstructs known structures with supertypes purely from the projected potential.
- This approach should allow us to rationalise peptide binding and TCR recognition and thus facilitate epitope identification and optimise transplantation strategies.

Abstract

Peptide-binding MHC proteins are thought the most variable proteins across the human population; the extreme MHC polymorphism observed is functionally important and results from constrained divergent evolution. MHCs have vital functions in immunology and homeostasis: cell surface MHC class I molecules report cell status to CD8+ T cells, NKT cells and NK cells, thus playing key roles in pathogen defence, as well as mediating smell recognition, mate choice, Adverse Drug Reactions, and transplantation rejection. MHC peptide specificity falls into several supertypes exhibiting commonality of binding. It seems likely that other supertypes exist relevant to other functions. Since comprehensive experimental characterization is intractable, structure-based bioinformatics is the only viable solution. We modelled functional MHC proteins by homology and used calculated Poisson-Boltzmann electrostatics projected from the top surface of the MHC as multi-dimensional descriptors, analysing them using state-of-the-art dimensionality reduction techniques and clustering algorithms. We were able to recover the 3 MHC loci as separate clusters and identify clear sub-groups within them, vindicating unequivocally our choice of both data representation and clustering strategy. We expect this approach to make a profound contribution to the study of MHC polymorphism and its functional consequences, and, by extension, other burgeoning structural systems, such as GPCRs.

Keywords

Major Histocompatibility Complex; Probabilistic Visualisation; HLA supertypes; Multi-level Gaussian Process Latent Variable Model.

1 INTRODUCTION

Innate and adaptive humoral and cellular immunity together form our ultimate bulwark against endemic infectious disease. Without the highly-organized, exquisitely-orchestrated immune response, our species, could not and would not have survived. The success of the immune response depends on identifying and responding to immunogenic protein antigens, achieved through the recognition of epitopes; in the context of cellular immunology, epitopes are short protein-derived peptides, recognized through the formation of ternary complexes comprising peptide, T-cell Receptor (TCR), and major histocompatibility complex (MHC). Recognition is coupled tightly to the peptide specificity of particular MHC alleles, and in turn the binding specificity of the TCR for an individual peptide-MHC complex (pMHC).

MHC proteins hold a key position in the immune system, playing fundamental roles in regulating immune responses, modulating the functional development of lymphocyte subsets, the acquisition and maintenance of self-tolerance, and the activation state and responses of host immune defenses. Cell surface MHC class I molecules report on the internal status of cells by presenting ligands for surveillance by CD8+ T cells, NKT cells and NK cells (Heinonen and Perreault, 2008). MHCs are poly-functional molecules involved *inter alia* with pathogen defense, homeostasis, smell recognition, and mate choice, play crucial clinical roles in transplantation rejection, have been implicated in mediating Adverse Drug Reactions, and have key roles in the nervous system.

On the basis of their biological properties and chemical structure, MHC proteins are grouped into two classes: class I and class II. MHC class I molecules typically present peptides from proteins synthesized within the cell (endogenous processing pathway). MHC class I proteins are encoded by three loci: HLA-A, HLA-B and HLA-C. MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway). MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ and HLA-DP.

The structure of the two classes differs. Class I MHC molecules comprise a transmembrane heavy chain of 44 kDa folded into a polymorphic peptide and TCR binding domain and a conserved immunoglobulin domain which binds CD8. Native class I MHCs form a complex with a small 12kDa protein called □2-microglobulin. Class II MHCs comprise 2 chains, □ and □, which both contribute to the peptide binding and TCR interaction site.

The peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α-chain) (Janeway *et al.,* 1999). Usually, only short peptides bind in extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only 9 amino acids actually occupy the site. Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif. The main prerequisite for a peptide to act as a T-cell epitope is that it binds to an MHC protein. The stable binding of the peptide to the MHC molecule is considered as the major bottleneck in the complicated pathway of antigen presentation. Class I MHC proteins bind peptides consisting of 8−11 amino acids, although longer peptides up to 15 amino acids in length are now starting to be identified, and originate from two sources: self-proteins and antigenic proteins. Their processing pathway involves the degradation of proteins by the proteasome, followed by transport mediated by the transporter associated with antigen processing (TAP) protein to the endoplasmic reticulum, where peptides are bound by MHC class I molecules. MHC−peptide complexes are then exocytosed to the cell surface, where they interact with TCRs expressed on the surface of T-cells.

Both MHC classes are polygenic (many genes) and extremely polymorphic (many alleles for each gene), increasing the probability that pathogens will contain many epitopes recognized within the whole population, which place restraints on a pathogen's capacity to escape immune control. More than 8,700 MHC molecules are listed in IMGT/HLA database (Robinson *et al.,* 2003). Peptide specificities across this vast array of MHCs are thought to form distinct clusters or supertypes (Sette and Sindney, 1999). Many studies, both experimental (Greenbaum *et al*., 2011) and computational (Harjanto *et al.,* 2014), have attempted to define such supertypes. Given the number and diversity of MHCS, the only tractable approach is a computational one. In this paper, we address the issue of extreme polymorphism in human class I MHC using informatics-driven theoretical methods. Building on our previous results (Doytchinova *et al*., 2004; Doytchinova and Flower, 2005), we construct 3-dimensional homology models of class I MHCs, from which we calculate Poisson-Boltzmann electrostatic potentials that act as descriptors for the application of novel state-of-the-art data reduction and visualisation methods.

2 METHODS

A set of protein sequences of HLA class-I were collected from the IMGT/HLA database (Robinson *et al*., 2003) (from July 2011 release for HLA-A, and from November 2011 release for HLA-B and HLA-C). The IMGT/HLA database nomenclature defines six parts of the HLA allele name. At first we excluded all those sequences which either have 'N' or 'L' or 'Q' as suffix at the end of the allele name. Secondly, from the rest of the allele set we have considered only those protein sequences that either have only one known DNA substitution within the coding region or if there is more than one DNA substitution, only the sequence with maximum length was considered. After excluding the sequences based on the explained criteria we selected 1,236 sequences of HLA-A, 1,779 of HLA-B and 929 of HLA-C. For structure-modeling purposes, a homology-modeling approach was used to model 3-dimensional structures in four steps using the Modeller software tool (Sali, 2010). We downloaded three known reference protein structures (i.e. HLA-A*0201 (`1I4F' protein data bank code) for the HLA-A, HLA-B*0801 ('1AGD' protein data bank code) for the HLA-B and HLA-CW*0401 ('1IM9' protein data bank code) for the HLA-C) retrieved from the protein data bank. The same three reference protein structures were previously used in (Doytchinova *et. al*., 2004). Selected sequences of each gene were aligned with the corresponding known reference protein structure. A few of the aligned sequences have shown some extra amino acids either at one or both ends since there was no match for them in the reference protein structure, so we optimized alignment by removing these segments to increase the similarity to the reference protein structure. All structures of HLA-B and HLA-C type were super-positioned on one of the structures of HLA-A based on the C-Alpha carbon atom. Side chain placement was performed using SCWRL (Bower *et al.*, 1997).

After structure modeling, the electrostatic potential (EP) was calculated in two steps: in the first step the transformation from the protein data bank (PDB) format to PQR format was performed using the software tool PDB2PQR (Dolinsky *et al.*, 2004,2007): this prepares structures for continuum electrostatic potential calculation by placing missing hydrogen atoms and any other heavy missing atoms. In the second step the Adaptive Poisson Boltzmann Solver (APBS) (Baker *et al*., 2001) was used to calculate electrostatic potentials by surrounding each protein structure with a three-dimensional grid box with 17^3 grid points (where the coarse grid covering the complete protein have lengths 210 angstrom (Â) in all three dimensions and a fine grid with 72, 32 and 52 angstrom (Â) in the x, y and z dimensions respectively focusing on the target area (i.e. whole area around the α1 and α2 regions)). Our interest is in analysing the top region (i.e. α1 and α2) of proteins and therefore we selected the $9*17^2=2601$ grid points which cover this region. Electrostatic potential outside the van der Waals surface is important for interactions with other molecules and therefore we ignored electrostatic potential at all points which were inside the van der Waals surface resulting in 2418 grid points which are outside the van der Waals surface of all the modeled

structures. So in our dataset each row represents a single protein structure and each column represents a grid position where the electrostatic potential was calculated.

3 RESULTS

In machine learning, transformation of high-dimensional data into a low dimensional space (usually 2D or 3D) is referred as data visualisation, data projection, or dimensionality reduction. Applying a linear data visualisation algorithm such as Principal Component Analysis (PCA) (Pearson, 1901), we observed no separation between three protein classes (as shown in Figure 1), indicating there may be value in using a non-linear data projection.

We therefore applied a state-of-the-art probabilistic non-linear data visualisation algorithm and found some interesting structure with clear separation between three protein classes (Figure 2). A brief theoretical description of probabilistic data visualisation is given in the following sub-section.

3.1 Probabilistic Visualisation

A latent variable model is used to represent a dataset $\mathbf{X} \in R^{N*D}$ with N data points in D-dimensions by mapping from a low-dimensional $\mathbf{Z} \in R^{N*Q}$ with N data points in Q-dimensions (usually $Q = 2 \; or \; Q = 3$ for visualisation). The mapping between a low-dimension data point $\boldsymbol{z}_n$ and a high-dimension data point $\mathbf{x}_n$ is defined by

$$x_{nd} = f_d(\mathbf{z}_n). \quad (1)$$

We use a random variable, τ_{nd}, to represent noise on the *d*th feature of the *n*th data point. Usually the noise model is Gaussian, so the conditional distribution of a data point $\mathbf{x}_n$ given a data $\mathbf{z}_n$ is

$$p(x_{nd}|\mathbf{z}_n) = f_d(\mathbf{z}_n) + \tau_{nd}, \quad (2)$$

If the mapping is assumed to be linear, $f_d(\mathbf{z}_n) = \mathrm{w}_d^T \mathbf{z}_n$, and the latent variable $\mathbf{z}$ is drawn from a Gaussian prior (with zero mean and unit variance), then the maximum likelihood solution of the model represents the principal subspaces of the data (Tipping and Bishop,1999): this is Probabilistic Principal Component Analysis (PPCA). Integrating out the latent variable gives the marginal likelihood

$$p(\mathbf{x}_n) = \int p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)\, d\mathbf{z}, \quad (3)$$

In a standard latent variable model, we optimize weights and marginalize out latent variables to maximize likelihood. Instead we use a variant of the latent variable model where weights are marginalized out and the latent space is optimized (i.e. the position of each data point is optimized). Such a model is known as the Gaussian Process Latent Variable Model (GPLVM) (Lawrence, 2004).

2.1 Gaussian Process Latent Variable Model (GPLVM)

In the GPLVM model, because weights are marginalized out, a conjugate prior over the weights is chosen to be a spherical Gaussian in each dimension

$$P(\mathbf{W}) = \prod_{d=1}^{D} N(\mathbf{w_i}|0, I). \qquad (5)$$

The likelihood after marginalizing out the weights is

$$\mathrm{P}(\mathbf{X}|\mathbf{Z}, \beta) = \prod_{d=1}^{D} p\left(\mathbf{x}_{(:,\boldsymbol{d})}|\mathbf{Z}, \beta\right), \qquad (6)$$

where $\boldsymbol{p}\left(\mathbf{x}_{(:,\boldsymbol{d})}|\boldsymbol{Z}, \boldsymbol{\beta}\right) = \boldsymbol{N}\left(\mathbf{x}_{(:,\boldsymbol{d})}\middle|\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}} + \boldsymbol{\beta}^{-\mathbf{1}}\boldsymbol{I}\right)$ represents a distribution over a single feature in the data space. GPLVM uses the following likelihood function as a criterion to optimise the latent variables (similar to the likelihood used in (Tipping and Bishop, 1999)).

$$\mathrm{L} = -\frac{DN}{2}\mathrm{Log}\,2\pi - \frac{D}{2}\mathrm{Log}\left(\det(\mathbf{K})\right) - \frac{1}{2}tr(\mathbf{K}^{-1}\mathbf{X}\mathbf{X}^{T}). \qquad (7)$$

If $\mathbf{K} = \mathbf{Z}\mathbf{Z}^{T} + \beta^{-1}I$ is a linear kernel then GPLVM is similar to PPCA. But for the GPLVM a non-linear RBF kernel is used (Lawrence, 2005). The optimization of the latent points can be achieved by first taking the gradient of the likelihood with respect to the kernel, and then combining this with, $\frac{\partial \mathbf{K}}{\partial z_{n,j}}$, using the chain rule. The gradient calculation uses the inverse of the kernel matrix; computing this has $O(N^3)$ complexity thereby making it less practical for large datasets. To migrate this sparse approximation is used which selects a small subset of the data points of size $k \ll N$ reducing the complexity from $O(N^3)$ to $O(k^2N)$. We used the GPLVM with an improved sparse approximation approach (Lawrence, 2008). This new improved approximation process was originally proposed for Gaussian process regression and is based on the unified view process as explained in (Quionero-candela *et al.*, 2005).

The standard GPLVM uses a mapping from the latent space to the data space for the training data only which constraints distant points in the data space to be distant in the latent space at the expense of local distance preservation. When users visualise data, it is the local structure that is most relevant to their analysis (for example, when they identify clusters). Therefore a variant of GPLVM which is constrained by a smooth back-projection (Lawrence, 2006) is employed to increase local neighborhood distance preservation because the data points $\boldsymbol{z}$ are no longer freely optimized. Instead they are the image of points $\mathbf{x}$ in the data space under the non-linear function like a Radial Basis function (RBF) kernel or multi-layer perceptron (MLP) kernel. This constrained mapping ensures that the data points which are close in the visualisation space are also close in the data space.

3.3 Quantitative Visualisation Quality Measures

Evaluating visualisation performance quantitatively is important but difficult because there is no known target output. We used quantitative quality measures that are based on the neighbourhood preservation to compare the mapping performance. These include trustworthiness, continuity (Venna and Kaski, 2001), mean relative rank errors with respect to data and visualisation space (Lee and Verleysen, 2008), Kullback-Leibler divergence (KLD) (Cover and Thomas, 1991) and nearest-neighbour classification error in the latent space as used in (Lawrence, 2005). A mapping is said to be *trustworthy* if k-neighbourhood in the visualised space matches that in the data space but if the k-neighbourhood in the data space matches that in the visualised space it maintains *continuity*.

For measuring the trustworthiness we consider $R_{i,j}^{\mathbf{X}}$ as the rank of the *j*th data point from the corresponding *i*th data point with respect to the distance measure in the high-dimensional data space $\mathbf{X}$, and $U_k(i)$ as the set of data points in the k-nearest neighbourhood of the *i*th data point in the latent space $\mathbf{Z}$ but not in k-nearest neighbourhood in the data space. The trustworthiness can be calculated with k-neighbours as

$$1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (R_{i,j}^{\mathbf{X}} - k). \qquad (8)$$

For measuring the continuity we consider $R_{i,j}^{\mathbf{Z}}$ as the rank of the *j*th data point from the corresponding *i*th data point with respect to the distance measure in the low-dimensional visualisation space $\mathbf{Z}$**,** and $V_k(i)$ as the set of data points in the k-nearest neighbourhood of the *i*th data point in the data space $\mathbf{X}$ but not in the k-nearest neighbourhood in the visualisation space. The continuity can be calculated with k-neighbours as

$$1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in V_k(i)} (R_{i,j}^{\mathbf{Z}} - k) \qquad (8)$$

Each mapping algorithm makes a tradeoff between trustworthiness and continuity: PCA often has higher continuity than trustworthiness whereas GPLVM with back constraints has higher trustworthiness than continuity as it focuses on preserving local distances (Venna and Kaski, 2005). We also use other quantitative quality measures such as mean relative rank errors with respect to data and latent space, which are similar to trustworthiness and continuity but use exact rank differences in the k-neighbourhood instead of the trustworthiness and continuity which only considers matches in the k-neighbourhood preservation ignoring the exact rank differences (Lee and Verleysen., 2008). For trustworthiness and continuity the higher the value (in the range 0 to 1) the better the proximity preservation whereas for mean relative rank errors, a lower value is better. In a visualisation context, trustworthiness and mean relative rank error with respect to data are more important than continuity as it ensures data points in k-neighbourhood of visualisation space are also neighbours in the data space (Kaski *et al.,* 2003) so any structure that is seen in the latent space is genuine. Table 1 shows that GPLVM performs better for our dataset than PCA.

We can analyze the performance in more detail. We prefer visualisations with a clear separation between gene classes: therefore we first apply mixture of Gaussians for each class and then compute Kullback-Leibler divergence (KLD). The higher the KLD measure the better the visualisation separates the classes. GPLVM gives much clearer separation between gene classes both by visual inspection and with respect to KLD measure (see Table 1). We also used nearest-neighbour classification error and, as shown in Table 1, GPLVM has a much lower nearest neighbour classification error than PCA confirming that the non-linear method outperformed the linear method.

3.4 Multi-Level GPLVM

We have developed a multi-level GPLVM (ML-GPLVM) algorithm to drill down into large-scale complex datasets to get more detailed insight (Mumtaz *et al*., 2014). The fundamental building block of our proposed ML-GPLVM approach is GPLVM with back-constraints using improved sparse approximation as discussed in section 3.2. In the ML-GPLVM framework, a root-level visualisation gives an overview of the complete dataset and a second/lower level visualisations give detailed views by building local visualisation models each trained using a subset of the dataset. To get a second level visualisation we can either apply automated clustering on the root level visualisation plot using techniques such as K-means (MacQueen, 1967) or Gaussian mixture models (GMM) (McLachlan and Basford, 1988), or allow the user to select clusters interactively by drawing polygons. These clusters are used to partition the dataset into subsets for training local visualisations. This process can be extended further in

each lower level plot to generate further levels. We present here an ML-GPLVM results using interactive clustering whereas detailed results using each of the clustering approach to generate local/detail views are available in (Mumtaz *et al*., 2014). We observed in an ML-GPLVM approach as we added visualisation levels, the results improved both visually (Figure 3) and quantitatively (Mumtaz *et al*., 2014).

3.5 Data Visualisation and Modelling System (DVMS)

We have included both GPLVM and ML-GPLVM in our visualisation tool: Data Visualisation and Modelling System (DVMS). DVMS can be freely downloaded from our research group website[1]. DVMS is based on the NETLAB machine-learning toolbox (Nabney, 2002) and is written in the Matlab programming environment. This visualisation tool facilitates user interaction with visualisation plots to generate tables of protein names or parallel coordinate plots of electrostatic potential values (Maniyar and Nabney, 2006). The user selects a region of interest either by drawing a containing polygon or by clicking on the region of interest to select nearest neighbours based on the Euclidean distance.

3.6 Comparison to Previous Supertype Analysis

Although our analysis is generated using a different means, it is instructive to compare our results briefly to previous work (Doytchinova *et al.,* 2004). It should be stressed that our results do not use any identifying characteristic of the MHC proteins other than their property distributions. Moreover, the present analysis differs from previous work, which focused on the peptide binding site only (Doytchinova *et al.,* 2004), so we should not expect an overwhelming overlap; yet all are sufficiently similar, and these commonalities between observed clusters are reassuring. With the exception of a few individual alleles, our current analysis effected a near complete separation of HLA-A, B, and C loci. Exceptions to this were clusters 8 and 12 (Figure 2), which may be indicative of some commonality of structural properties corresponding to convergent evolution of HLA-A and HLA-B alleles. Even here, alleles were almost completely separated but in a continuous distribution that is hard to separate further without prior knowledge. It is interesting that HLA-C forms six well separated clusters, which contrasts sharply with previous results, and is perhaps suggestive of the greater variability in the extended surface analyzed here relative to HLA-A and B. This is functionally consistent with the interactions made by HLA-C with a wider range of receptors other simply TCRs.

By rigorous state-of-the-art analysis of projected properties, we have identified clusters corresponding to the three class I human MHC loci, and sub-groups therein. It is notable that the analysis recovers the HLA-A, HLA-B, and HLA-C alleles using only their property distributions, without prior knowledge of a division by loci; information used only when labelling the result plots. This gives confidence to any assertion we might make regarding the division of the allele population into structurally and functionally similar sub-groups. This

[1] http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/downloads/

information will inform accurate identification of T-cell epitopes, a crucial step when developing epitope ensemble vaccines.

The three different class I HLA loci are possessed of functional differences, such as binding NK receptors, system differences, such as the breadth of anti-HIV responses of different HLA loci (Kiepiela *et al*., 2004), as well as structural ones, including the observation that different loci have peptide repertoires that are distinct in their size and specificity (Paul *et al.,* 2013). Thus our ability to distinguish the three loci so unequivocally is notable. It implies that the differences are sufficiently strong to be obvious at the level of projected properties alone, and this gives credence to our identification of further subsets within the individual loci.

5. DISCUSSION

The vertebrate MHC exhibits extreme, even unprecedented, polygenic genetic diversity or polymorphism. MHC diversity is believed to strongly influence host susceptibility to infectious, autoimmune, and other diseases. MHC diversity at all loci has evolved under selection pressure from a diverse and variable pathogen population. MHC supertypes are thought to arise through both divergent and convergent molecular evolution. Such processes will concentrate acceptable mutations at certain sequence positions where conflicting constraints are best balanced. It is thought that certain alleles offer protection against particular pathogens, as they preferentially bind conserved epitopes characteristic of prevalent pathogens. Thus geography strongly influences allelic diversity, since pathogen distributions are typically constrained in space as well as time. Mechanistically, the high diversity of human MHCs and thus MHC binding specificities is driven primarily by the high sequence diversity of short epitopes within pathogenic proteins presented for immune surveillance. The principal rationale for MHC evolution is thus antagonistic co-evolution (or Red Queen hypothesis), where pathogens evolve to evade immune surveillance by common MHC alleles, providing a selective advantage to hosts with rare MHC alleles; this cyclic process necessarily leads to high MHC diversity. The MHC also evolves under several other selection pressures, including mate and foeto-maternal selection, smell reception, *etc*. The resulting MHC polymorphism is seemingly unmatched by other human proteins and represents a considerable challenge to traditional experimentation.

The results of our analysis are fully consistent with both the choice of Poisson-Boltzmann electrostatic potential as a meaningful surrogate of all through-space molecular interactions and with the sophisticated methods of data reduction used to view and derive the final clustering. They are also consistent with the foregoing evolutionary argument, since they suggest that, with the exception of a handful of genes, the three class I loci exhibit quite distinct specificities for peptides and TCRs. Redundant specificities shared between loci would be not favourable since this would reduce the diversity of peptides a host could recognize and respond to, and likewise reduce the diversity of pathogens the host immune system could

effectively combat. It will be interesting to extend our analysis to investigate the structural basis for this phenomenon.

To be recognized by the immune system, peptides must be presented by an MHC on the surface of an antigen presenting cell (APC) for inspection by T-cells. Not all presented peptides are recognized by T-cells. Those recognized are known as epitopes. A T-cell epitope is a continuous peptide of variable length. Peptide vaccines target epitopes against which a protective response can be induced. The immune system expresses many different TCRs, each with its own fine specificity for pMHC's. The repertoire of dominant and nondominant TCR specificities is largely shaped by the infection and vaccination history of an individual. The number of TCRs present in appreciable quantities will still be many orders of magnitude greater than the three different class I MHCs. Usually, relatively little is known about the repertoire of TCR specificities. Nonetheless, the pivotal event in epitope recognition is the formation of the ternary pMHC complex.

The binding of TCRs to pMHC is relatively weak (micromolar) compared to peptide binding to the MHC (nanomolar). Thus, peptide binding is a necessary, but not a wholly sufficient condition, for a peptide to be an epitope. It is now accepted that workable T-cell epitope prediction methods depend on the accurate prediction of the most discriminatory step of the antigen presentation process: peptide binding to MHCs. Unfortunately, MHC polymorphism means that only a small proportion of alleles have had the peptide specificity properly characterized. Our current analysis is a potential route to circumvent this issue, complementary to pocket stitching and molecular dynamics as approaches, allowing the specificity within clusters to be inferred. Thus for a newly identified or otherwise uncharacterized allele, its specificity could be approximated using its sequence to suggest the closest known specificity, followed by the specific identification of potential binders to this allele, in-turn followed by experimental verification of any suggested epitopes. This is feasible since the GPLVM and ML-GPLVM methods can be used to project previously unseen data into the latent space and thus enable the user to see how the new allele relates to better understood alleles.

Our analysis has targeted the top face of the MHC, including the peptide-binding site, where the TCR is known to bind from crystal structures (Bhati *et al*., 2014). Thus our analysis is also germane to discussion of transplantation rejection. While the causes of chronic transplant rejection are complex, and poorly understood, and hyper-acute rejection is mediated by antibodies and complement, T-cells are implicated as the prime arbiter of rejection for accelerated and acute rejection. This occurs through direct (response to peptides bound by donor MHC), indirect (response to non-self-peptides bound by recipient MHC), and semi-indirect routes (mixture of above). Thus HLA matching is thought to crucial to the success of transplantation. Matching is usually expressed in terms of numbers of identical loci matched rather than the similarity of non-identical alleles. MHC-centric yet empirical methods such as

HLAmatchmaker (Duquesnoy, 2011), go part way to addressing this. Several studies have addressed the same issue but from a truly 3-dimensional perspective, using structural models to evaluate the effect of HLA-mismatches on peptide binding (Dudkiewicz *et al*., 2009; Yanover *et al*., 2011). Thus our results analysis may be seen as a useful extension to such studies and complementary to this approach.

As the verity of *in silico* predictions improves, so subsequent experimental work should become more efficient and successful. We will seek to extend our approach, using it to classify MHC alleles in terms of peptide specificity, TCR specificity, and antibody interaction and use it to investigate practical problems in epitope prediction, solid organ and bone marrow transplantation, mate-choice, and MHC-mediated adverse drug reactions.
The present approach, which combines the established protocol of target family landscape profiling (Naumann and Matter, 2002; Doytchinova *et al.,* 2004; Doytchinova and Flower, 2005; Pirard and Matter, 2006), with a novel means of calculating projected properties and state-of-the-art data visualisation and clustering, is clearly promising. Each MHC allele has a unique sequence, and thus unique 3-dimensional structure and functional properties - in our case their binding specificity for peptides and TCRs - and the present study has compared MHC structure as a means to classify MHCs in terms of such interactions. We expect that this techniques, and developments thereof, will make a profound contribution to the study of MHC polymorphism and its functional consequences, and, by extension, other burgeoning structural systems, such as G-Protein Coupled Receptors (Bjarnadóttir *et al*., 2006) and Kinases (Endicott and Noble, 2013).

Funding: Faculty development program of the Islamia University of Bahawalpur (IUB), Pakistan.

ACKNOWLEDGEMENTS

Shahzad Mumtaz thanks Prof. Belal A. Khan (Ex Vice Chancellor of the IUB, Pakistan) for arranging funding of his PhD.

REFERENCES

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA 98, 10037-10041.

Bhati M, Cole DK, McCluskey J, Sewell AK, Rossjohn J. (2014) The versatility of the αβ T-cell antigen receptor. Protein Sci. 23:260-272.

Bjarnadóttir TK, Gloriam DE, Hellstrand SH, Kristiansson H, Fredriksson R, Schiöth HB(2006) . Comprehensive repertoire and phylogenetic analysis of the Gprotein-coupled receptors in human and mouse. Genomics. 88:263-273.

Bower, M. J., Cohen, F. E., and Dunbrack, J. (1997) Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modelling. J. Mol. Biol, 267:1268–1282.

Cover, T. M., Thomas, J. A. (1991) Elements of information theo-ry. New York: Wiley. ISBN: 0471062596 9780471062592

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA. (2007) PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res, 35, W522-5.

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. (2004) PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res, 32, W665-W667.

Doytchinova, I. A., Guan, P., and Flower, D. R. (2004) Identify-ing human MHC supertypes using bioinformatics methods. The Journal of Immunology, 172:4314–4323.

Doytchinova IA, Flower DR. (2005) In silico identification of supertypes for class II MHCs. J Immunol. 174:7085-7095.

Dudkiewicz M, Malanowski P, Czerwin Ski J, Pawłowski K. (2009) An approach to predicting hematopoietic stem cell transplantation outcome using HLA-mismatch information mapped on protein structure data. Biol Blood Marrow Trans-plant. 15:1014-10125.

Duquesnoy RJ. (2011) Antibody-reactive epitope determination with HLAMatchmaker and its clinical applications. Tissue Antigens. 77:525-534.

Endicott JA, Noble ME (2013). Structural characterization of the cyclin-dependent protein kinase family. Biochem Soc Trans. 41:1008-1016.

Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. (2011) Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. Immunogenetics. 63:325-335.

Harjanto S, Ng LF, Tong JC. (2014) Clustering HLA Class I Superfamilies Using Structural Interaction Patterns. PLoS One. 9:e86655.

Heinonen KM, Perreault C. (2008) Development and functional properties of thymic and extrathymic T lymphocytes. Crit Rev Immunol 28: 441–466.

Janeway CA, Walport TM, Capra JD. (1999) Immunobiology: the immune system in health and disease. Current Biology Publi-cation, London,

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P. & Castrén, E. (2003). Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics, 4, 48.

Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, Rathnavalu P, Moore C, Pfafferott KJ, Hilton L, Zimbwa P, Moore S, Allen T, Brander C, Addo MM, Altfeld M, James I, Mallal S, Bunce M, Barber LD, Szinger J, Day C, Klenerman P, Mullins J, Korber B, Coovadia HM, Walker BD, Goulder PJ. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature. 2004 432:769-775.

Lawrence, N. D. (2005) Probabilistic Non-linear Principal Com-ponent Analysis with Gaussian Process Latent Variable Models. Journal of Machine Learning Research, 6, 1783-1816.

Lawrence, N. D. & Candela, J. Q. (2006) Local distance preservation in the GPLVM through back constraints. In W. W. Cohen & A. Moore (eds.), ICML (p./pp. 513-520), : ACM. ISBN: 1-59593-383-2

Lawrence, N. D. (2008) Large scale learning with the Gaussian process latent variable model. Technical report, university of sheffield, United Kingdom.

Lee, J. A. & Verleysen, M. (2008) Rank-based quality assessment of nonlinear dimensionality reduction. ESANN (p./pp. 49-54) .

Maniyar, D. M. & Nabney, I. T. (2006). Visual data mining using principled projection algorithms and information visualization techniques.. In T. Eliassi-Rad, L. H. Ungar, M. Craven & D. Gunopulos (eds.),KDD (p./pp. 643-648), : ACM. ISBN: 1-59593-339-5

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1, pages 281–297. University of California Press, Berkeley, CA, USA.

McLachlan, G. and Basford, K. (1988) Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York

Mumtaz, S. Nabney, I.T and Flower, D.R. (2014) Multi-level visualization using Gaussian Process Latent Variable. In Proceed-ings of the 5th International conference Information Visualization Theory and Applications (IVAPP), pages 122-129, Lisbon Portugal, SCITPRESS

Nabney, I. T. (2002). NETLAB. Algorithms for Pattern Recognition. Springer. ISBN: 1852334401

Naumann T, Matter H. (2002). Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. J Med Chem. 2002 45:2366-2378.

Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. 2013 HLA class I alleles are associated with peptide binding repertoires of different size, affinity, and immunogenicity. J Immunol.;191(12):5831-9.

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space, Philosophical Magazine 2: 559-572.

Pirard B, Matter H (2006). Matrix metalloproteinase target family landscape: a chemometrical approach to ligand selectivity based on protein binding site analysis. J Med Chem. 49:51-69.

Quionero-candela, J., Rasmussen, C. E., and Herbrich, R. (2005) A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research, 6:1939–1959.

Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SGE. (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex, Nucleic Acids Research, 31:311-314

Sali, A. (2010) MODELLER: A Program for Protein Structure Modelling Release 9v8 r7145. http://www.salilab.org/modeller/.

Sette A, Sidney J. (1999) Nine major HLA class I supertypes ac-count for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics. 50:201-212.

Tipping, M. E., and Bishop C. M. (1999) Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 61, No.3, pp. 611–622.

Venna, J. and Kaski, S. (2001) Neighbourhood preservation in non-linear projection methods: An experimental study. In Proceed-ings of the International Conference on Artificial Neural Net-works, ICANN ’01, pages 485–491, London, UK, Springer-Verlag.

Venna, J., & Kaski, S. (2005) Local multidimensional scaling with controlled trade-off between trustworthiness and continuity. http://eprints.pascal-network.org/archive/00001233/

Yanover C, Petersdorf EW, Malkki M, Gooley T, Spellman S, Velardi A, Bardy P, Madrigal A, Bignon JD, Bradley P. (2011) HLA mismatches and hematopoietic cell transplantation: structural simulations assess the impact of changes in pep-tide binding specificity on transplant outcome. Immunome Res.7:4.

Figure 1: PCA visualisation of HLA class 1 dataset (cyan dots ('.') for HLA-A, red plus sign ('+') for HLA-B and blue squares ('□') for HLA-C)

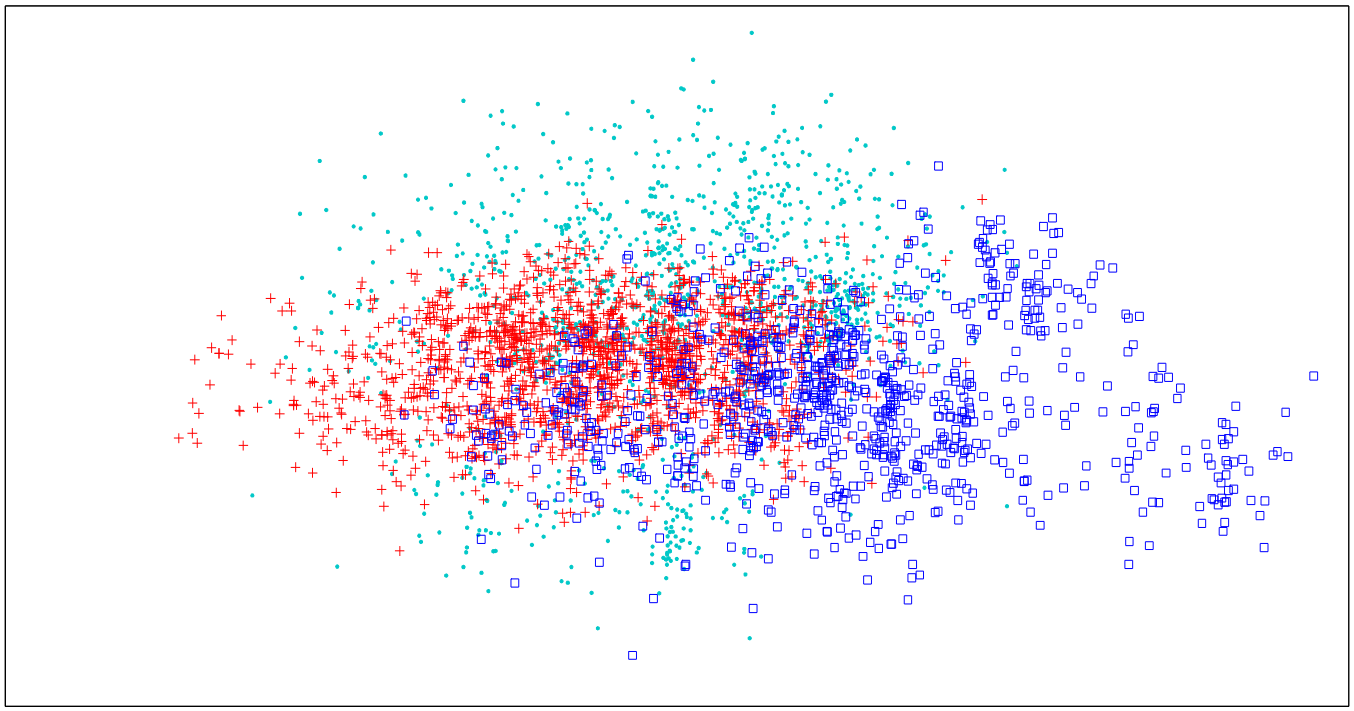
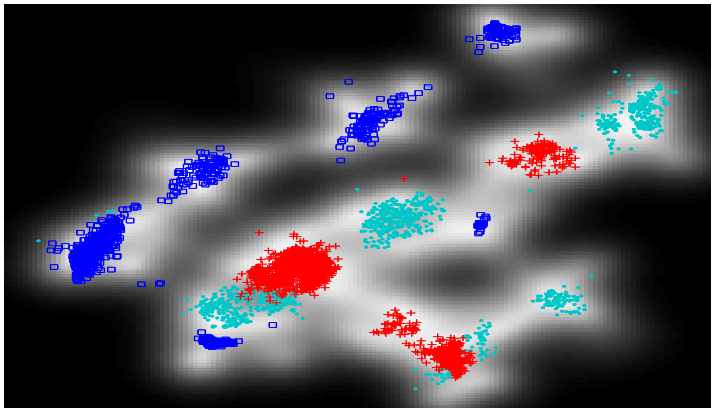
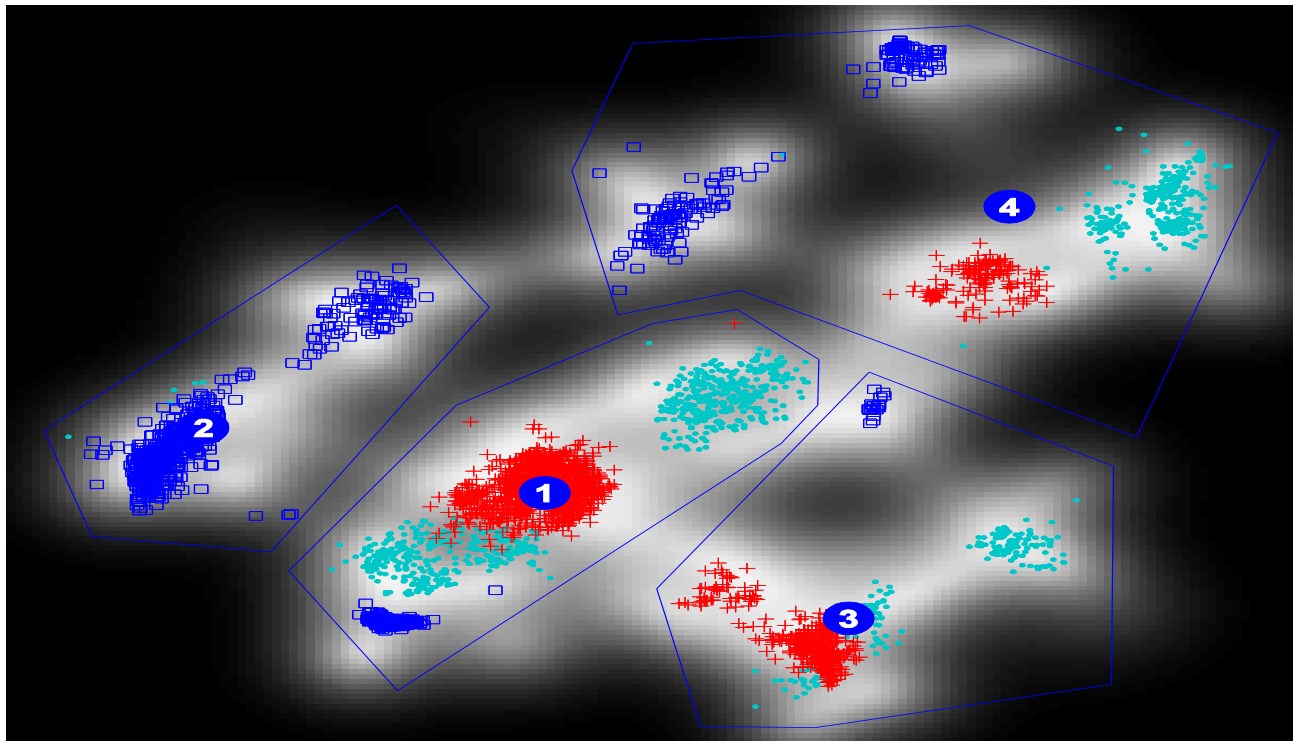
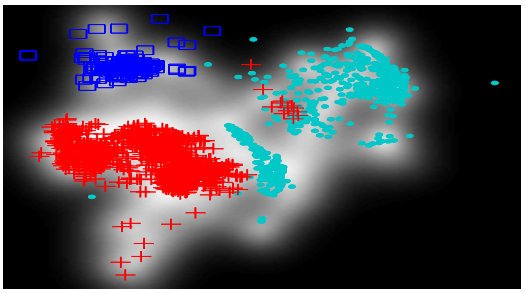


Figure 2: GPLVM visualisation (cyan dots ('.') for HLA-A, red plus sign ('+') for HLA-B and blue squares ('□') for HLA-C). The grey background shows the mapping precision (lighter regions have better precision.).

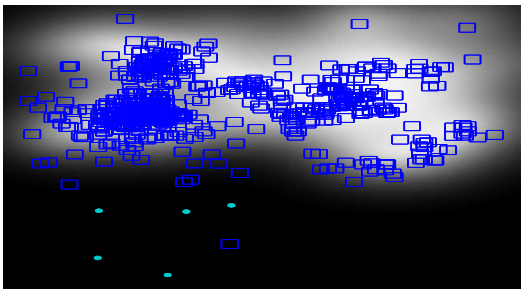




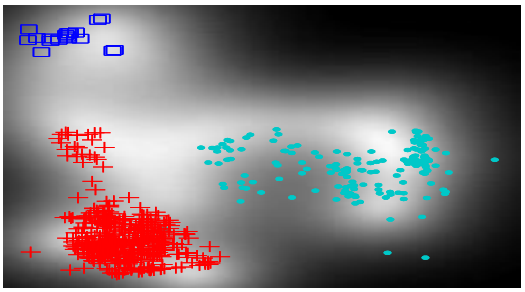
Submodel 1



Submodel 2



Submodel 3



Submodel 4

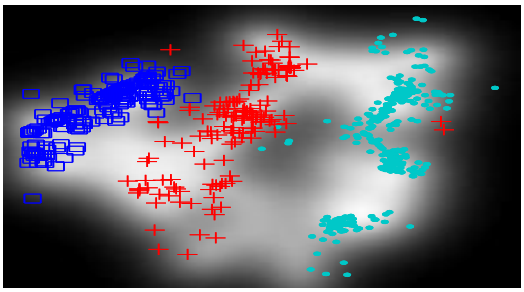


Figure 3: ML-GPLVM visualisation of 'MHC class-I' dataset with interactive clustering. (a) Root level visualisation plot (numbered blue circles with numbers indicate cluster centres and blue

lines represent cluster boundaries and data points and background as defined in Figure 2. (b) Level-2 visualisation plots.

Table 1: Quantitative quality comparison of linear and non-linear visualisation of HLA Class I dataset.

	PCA	**GPLVM**
Trustworthiness	0.9137	**0.9707**
Continuity	**0.9666**	0.9069
MRREdata	0.0082	**0.0078**
MRRElatent	**0.0068**	0.0076
Kullback-Leibler divergence (KLD)	16.6685	**152.6811**
Nearest Neighbour Error (%)	35.72	**1.52**