Bioinformatics, 32(15), 2016, 2297–2305
doi: 10.1093/bioinformatics/btw156
Advance Access Publication Date: 29 March 2016
Original Paper

Structural bioinformatics

HLaffy: estimating peptide affinities for Class-1 HLA molecules by learning position-specific pair potentials

Sumanta Mukherjee[1], Chiranjib Bhattacharyya[2] and Nagasuma Chandra[3,*]

[1]IISc Mathematics Initiative, Indian Institute of Science, Bangalore, India, [2]Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India and [3]Department of Biochemistry, Indian Institute of Science, Bangalore, India

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on September 30, 2015; revised on February 12, 2016; accepted on March 3, 2016

Abstract

Motivation: T-cell epitopes serve as molecular keys to initiate adaptive immune responses. Identification of T-cell epitopes is also a key step in rational vaccine design. Most available methods are driven by informatics and are critically dependent on experimentally obtained training data. Analysis of a training set from Immune Epitope Database (IEDB) for several alleles indicates that the sampling of the peptide space is extremely sparse covering a tiny fraction of the possible nonamer space, and also heavily skewed, thus restricting the range of epitope prediction.

Results: We present a new epitope prediction method that has four distinct computational modules: (i) structural modelling, estimating statistical pair-potentials and constraint derivation, (ii) implicit modelling and interaction profiling, (iii) feature representation and binding affinity prediction and (iv) use of graphical models to extract peptide sequence signatures to predict epitopes for HLA class I alleles.

Conclusions: HLaffy is a novel and efficient epitope prediction method that predicts epitopes for any Class-1 HLA allele, by estimating the binding strengths of peptide-HLA complexes which is achieved through learning pair-potentials important for peptide binding. It relies on the strength of the mechanistic understanding of peptide-HLA recognition and provides an estimate of the total ligand space for each allele. The performance of HLaffy is seen to be superior to the currently available methods.

Availability and implementation: The method is made accessible through a webserver http://proline.biochem.iisc.ernet.in/HLaffy.

Contact: nchandra@biochem.iisc.ernet.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recognition of short peptides by HLA molecules is a fundamental requirement for triggering T-cell responses. HLA genes are highly polymorphic forming hundreds of different alleles (Sette and Sidney, 1999). The alleles understandably have a common scaffold and exhibit variations in their peptide binding pockets (Murphy *et al.*, 2007) so that each allele preferentially binds a distinct subset of peptides. In an individual, cells expressing up to six different alleles are present, each capable of recognizing a large number of peptides. Such multi-specificity enables them to recognize a wide array of

potential antigens from pathogenic organisms and other sources. The most preferred length of the peptide is 9 amino acids, although peptides with 8–10 amino acids, are also recognized in some cases (Hewitt, 2003). The number of possible nonameric peptides that are theoretically possible is as high as 512 billion. Experimental identification of such a large repertoire of epitopes for each allele is virtually impossible, making it important to have reliable and accurate methods for predicting epitopes for different alleles. The need for powerful methods to predict potential epitopes has become even more important as hundreds of bacterial and viral genomes have been sequenced, increasing the need for understanding their antigenicity profiles.

A number of methods have been developed for predicting T cell epitopes over the last two decades (Lundegaard *et al.*, 2010). They can be broadly divided into four classes, as methods based on: (i) sequence motif detection, (ii) empirical scoring matrices and profiles implicitly encoding more complex sequence patterns, (iii) machine learning algorithms applied to experimentally determined data and (iv) structure-based methods. The motif methods are the first to be developed among all the prediction methods and are based on the identification of preferred amino acid residues at different positions. Some positions which were identified as heavily contributing to recognition were termed 'anchor residues' in the nonameric peptides (Hobohm and Meyerhans, 1993). The motifs or patterns describing the positions of the anchors and the residue preferences were derived for different alleles (Rammensee *et al.*, 1999). Anchor motifs have been used to search for antigenic regions in sequences of infectious viruses, bacteria and parasites. The amino acid preferences in these motifs however turn out to be an insufficient criteria for recognition, leading to prediction accuracies of only about 70%. Although anchor residues contribute significantly to the HLA-peptide binding, large scale analysis of experimental data have shown that other residue positions on binder peptides contribute equally in stabilizing the peptide-HLA complex. Consequently new scoring matrices were derived by analyzing experimentally measured binding affinities for hundreds of different peptides to each allele (Parker *et al.*, 1994). However, these methods assume that the contribution of different residues were independent of each other and most models were additive in nature.

Data-driven methods were developed to utilize available large-scale experimental data. One such approach is regression analysis that estimates the response function, which in this case is binding affinity, from a given set of data and response variables, generally using a polynomial function. Regression analysis based methods were used to obtain quantitative prediction in terms of half lives of dissociation of β-microglobulin from the HLA complex. Other machine learning (Lundegaard *et al.*, 2008; Peters and Sette, 2005; Roomp *et al.*, 2010) and QSAR approaches (Doytchinova *et al.*, 2004) that have been attempted include artificial neural networks and hidden Markov models that correlate measurable properties of different peptides with their binding potential.

Approaches that use explicit structural information in general are believed to provide detailed insights into the process of peptide recognition, thereby resulting in more accurate predictions. However, progress in this direction has been very slow, primarily because structural modelling is a computationally intensive process, rendering it feasible only on a small scale (Kumar and Mohanty, 2007; Schueler-Furman *et al.*, 2000). More recently, a method based on molecular dynamics simulations and estimation of free energy of binding between peptide and HLA molecules has been attempted (Yanover and Bradley, 2011). The applicability of this method is seen to be limited due to the amount of time each simulation requires. Methods that indirectly capture structural information have also been used, which show reasonable prediction accuracies (Hoof *et al.*, 2009; Lundegaard *et al.*, 2008).

This work introduces a new method HLaffy, for prediction of peptide HLA binding affinity. The method stands on the strength of a mechanistic model of peptide-HLA recognition. Distinct advantages of this method are: (i) it uses structural information explicitly, (ii) it samples the naturally occurring peptide space in bacteria and viruses systematically and hence is not restricted by available experimental data, (iii) it assigns weights to different interactions and through that, it identifies important pair potentials and (iv) it uses a graphical model to represent the epitope pool efficiently. The method is implemented on a webserver, made accessible at http://proline.biochem.iisc.ernet.in/HLaffy/.

2 Methods

2.1 Datasets

The list of 2010 HLA class-1 alleles was obtained from ImMunoGeneTics database (IMGT) (Robinson *et al.*, 2013). For 43 of these alleles, information about peptides they recognize along with their individual binding strengths, was obtained from *IEDB* (Vita *et al.*, 2010). Similar but independently curated data for HLA-peptide association for 32 alleles is available from MHC Binding, Non-binding peptide database (*MHCBN*) (Bhasin *et al.*, 2003). Only those unique to *MHCBN* were taken for validation purposes. Crystal structures of 21 different alleles as complexed with their respective peptides are available from PDB (Berman *et al.*, 2000).

2.2 Implementation and web-server

Affinity estimation and peptide ranking modules are computationally intensive, and are therefore implemented in C++. Publicly available GLPK library is used for solving linear optimization. Eigen3 library is used for linear algebraic operations and solving the graphical model (http://eigen.tuxfamily.org). The prediction tool is made publicly available at http://proline.biochem.iisc.ernet.in/HLaffy.

The HLaffy suite involves four core modules. Figure 1 shows the different computational steps involved in these processes and their logical dependencies. (i) Structural modelling of $\approx$ 16 000 complexes from a dataset of experimentally determined peptides, followed by computation of statistical pair-potentials. (ii) Design of a representative peptide library and obtaining allele specific inter-molecular residue–residue contact profiles. (iii) Implicit modelling of peptide binding modes and feature extraction and binding affinity estimation. (iv) Deriving a graphical model for each allele, and obtaining a sequence profile of the predicted set of high-affinity epitopes.

2.3 Structural modelling of peptide-HLA complexes

This computational module includes three distinct steps: (i) obtaining a dataset of known peptides, (ii) molecular modelling and (iii) computing statistical contact potentials. First, a set of $\approx$ 16 000 experimentally known peptides with respective HLA alleles was obtained from the *IEDB* database. Peptides with binding affinities $\leq$ 500 nM are considered to be strong binders. Models of HLA-molecules are built using homology modelling (Sali and Blundell, 1993). In Class-1 HLA molecules, the peptide binding groove is closed at both ends and narrow, thereby restricting conformational flexibility for the bound peptides. Structural superposition of nonameric bound peptides obtained from the PDB shows high

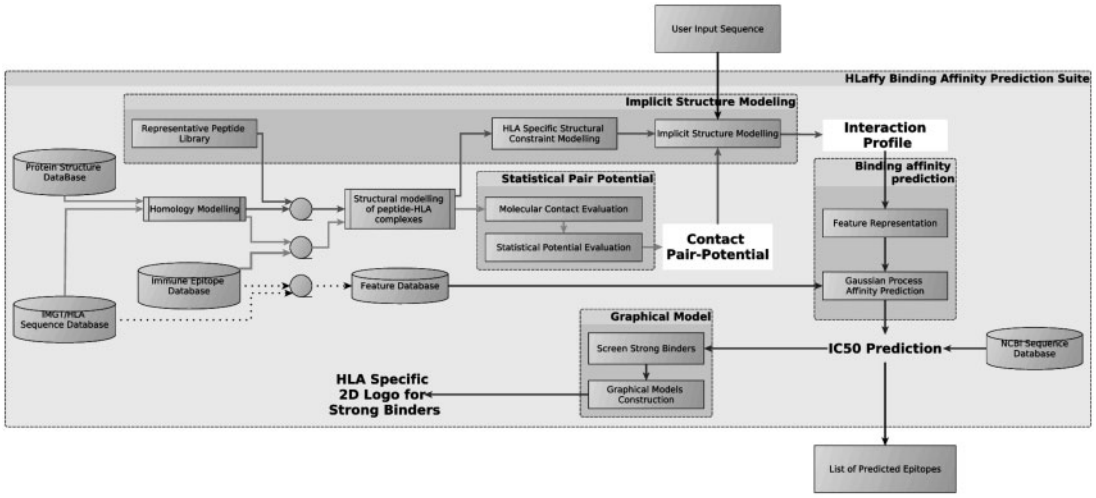

Fig. 1. The HLaffy workflow illustrating four main modules which comprises the prediction suite and their logical dependencies. (**a**) Structural modelling, (**b**) computing contact pair potentials, (**c**) interaction profiling and (**d**) graphical models for epitope pool representation. An user can input a query genome or individual protein sequence and obtain a set of predicted epitopes as the output. External resources used in the prediction suite are also indicated

conservation in the backbone structures, all in an extended conformation (≤ 1 Å). Hence, the peptide complexes are built by obtaining the peptide backbone from the template structures and modelling the side chains by a standard rotamer search. SCWRL4 (Krivov *et al.*, 2009) software was used for fixing rotamers. Complexes thus obtained were further energy minimized using GROMACS (Hess *et al.*, 2008; Van Der Spoel *et al.*, 2005). This process produces atomic level structural models of peptide-HLA complexes, allowing a detailed residue contact analysis.

2.4 Statistical pair-potentials for peptide HLA interaction interfaces

Statistical potentials capture information bias in a knowledge base. If one occurrence of one event influences the occurrence of another, they tend to show some extent of correlation in their joint occurrence order. This particular fact is captured through statistical potential. In this context, events are occurrences of interactions between a peptide residue and a HLA residue at the binding site. The evaluation of the statistical pair potential, also referred to as contact pair potential (*CPP*) from a given dataset is given as

$$u(a_i, a_j) = -kT \cdot \log\left(\frac{N_{\text{obs}}(a_i, a_j)}{N_{\text{exp}}(a_i, a_j)}\right) \quad (1)$$

where N_{exp} and N_{obs} are the expected and observed number of interactions between a peptide and a HLA residue of types a_i and a_j. $u(a_i, a_j)$ is the contact potential between them. k and T represent the Boltzmann constant and temperature respectively. The estimation of statistical potential (Eq. 1) demands accurate evaluation of residue-residue contacts (N_{exp}) between the peptide and the HLA. Non-covalent interactions between two residues are decided by their chemical nature and atomic distances between them. The polar nature of the solvent causes a local screening effect, due to which the interaction strength dies off quickly with increasing distance. The varying side chain length and flexibility results in a different influence radius for different amino acids. To determine expected contacts (N_{exp}), first the feasible dihedral sampling space was estimated through equilibrium molecular dynamics simulations of alanine flanked tri-peptides in water (for 2 ns). The dihedral space is represented as a solid cone encompassing a conformational space cone for each residue type. The smallest amino acid glycine does not have a side-chain and therefore the influence area for glycine is described using a sphere, while, for other residues, the influence volume is represented as a cone. The cone is described using two parameters, the solid angle subtended at C_α, with its axis defined by the vector joining $C_\alpha - C_\beta$ atoms of the residue, and the height of the cone. In a peptide-HLA complex, if the conformational space cones of any two residues intersect, they are considered to have the potential to interact and therefore included in the 'expected interactions' count. The entire set of residue pairs whose cones intersect contribute to the N_{exp} parameter in Eq. 1. If a peptide residue lies within a distance of 4.5 Å of the HLA site residue, they are said to be in contact and counted for the N_{obs} parameter. Estimations for N_{exp} and N_{obs} are described below

$$N_{\text{obs}}(a_i, a_j) = \sum_{k=1}^{N} \sum_{u=1}^{l_p} \sum_{v=1}^{l_h} \delta\left(\alpha\left(p_u^k\right) = a_i, \alpha\left(h_v^k\right) = a_j\right) C_{\text{obs}}\left(p_u^k, h_v^k\right)$$

$$N_{\text{exp}}(a_i, a_j) = \sum_{k=1}^{N} \sum_{u=1}^{l_p} \sum_{v=1}^{l_h} \frac{\delta\left(\alpha\left(p_u^k\right) = a_i, \alpha\left(h_v^k\right) = a_j\right)}{|\text{neighbor}\left(p_u^k\right)|} N_{\text{obs}}\left(p_u^k, h_v^k\right)$$

where $C_{\text{obs}}(p_u^k, h_v^k)$ is the number of contacts observed in the structure k between peptide residue p_u and HLA residue h_v. $\alpha(\bullet)$ maps the peptide or HLA residue position to the corresponding amino-acid; neighbor(p_u^k), returns neighbouring HLA residues around p_u, in structure k, that is inside the possible interaction sphere.

2.5 Designing a representative peptide library (RPL)

There are over 20 naturally occurring amino-acids. Peptides are short polymers of amino-acids. Typical MHC binding peptide ligands are nine-residue long, which theoretically leads to 20^9 different peptides. This space is too large for a thorough exploration. Hence a small representative library of peptides has been designed, such that the peptide space is uniformly sampled. The designed library ensures that all possible binary tuples have been sampled equally for any two positions of the design nonameric peptide sequences. For a combination of any two positions i and j, where $i, j \in \{1 \ldots 9\}$ and $i \neq j$, all 400 possible tuples are sampled.

The design of this library involves cyclic codes. All possible tuples (pair) generated from a character set (20 amino acids), can be

factorized into a set of circular permutation groups. Groups are abstract algebraic structures. It refers to a finite/infinite set of elements that satisfies closure, associativity, identity property and inverse property. Permutation groups consist of elements corresponding to permutations of all elements of a given set M. Permutation groups are written as

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \dots & \sigma(n) \end{pmatrix} \quad (2)$$

where $\sigma(i)$ is a corresponding permutation map of position i. A circular permutation generates the element permutation map as $\sigma(i) = (i + k) \bmod n$, denoted by $\mathbb{P}^{o}_{n,k}$, where k is a number between 0 to n, n being the number of elements the set contains. 0 refers to an identity permutation. Due to the group property $\mathbb{P}^{o}_{n,k_i} \times \mathbb{P}^{o}_{n,k_j} \in \mathbb{P}^{o}_{n,*}$. All possible tuples can be factored as $S \times S = \cup_{k=0\dots|S|} \{S \bullet \mathbb{P}^{\circ}_{|S|,k}\}$. The $\bullet$ operator defines concatenation of two fields, generating a tuple. It must be noted that, in case of a prime value of n, the tuple set generated by this process is non-overlapping and covers the entire tuple set. Therefore, $S \bullet \mathbb{P}^{\circ}_{n,k_1} \bullet \mathbb{P}^{\circ}_{n,k_2} \bullet \dots \mathbb{P}^{\circ}_{n,k_p}$, defines a $(p+1)$-length peptide sequence. Right from the very first sequence position, for any two positions i and j in the sequence, the sequence permutation can be given as $P^{\circ}_{|S|+k_j-k_i}$. In order to efficiently explore the peptide nonameric sequence space, character sets were extended to prime number cardinality by repeating naturally frequently occurring amino acid residues. The peptide library generator sequence is then given as

$$\begin{array}{cccccccc} \mathbb{P}^{\circ}_{n,0} & \bullet & \mathbb{P}^{\circ}_{n,1} & \bullet & \mathbb{P}^{\circ}_{n,2} & \bullet \dots \bullet & \mathbb{P}^{\circ}_{n,p} \\ \mathbb{P}^{\circ}_{n,1} & \bullet & \mathbb{P}^{\circ}_{n,3} & \bullet & \mathbb{P}^{\circ}_{n,5} & \bullet \dots \bullet & \mathbb{P}^{\circ}_{n,(2p+1) \bmod n} \\ \vdots & & \vdots & & \vdots & \vdots & \vdots \\ \mathbb{P}^{\circ}_{n,n-2} & \bullet & \mathbb{P}^{\circ}_{n,(2n-4) \bmod n} & \bullet & \mathbb{P}^{\circ}_{n,5n-6} & \bullet \dots \bullet & \mathbb{P}^{\circ}_{n,((p+1)n-(2p+2)) \bmod n} \end{array} \quad (3)$$

The design steps are explained in Algorithm 1. 'A' refers to set of all possible amino-acids. The design reports a matrix of size $m^2 \times (m+1)$, m being the size of the character set used for the design. As per this design, all possible amino acid pairs between any two positions are explored. This design strategy can generate a peptide library of maximum sequence length m which is same as the length of the character set. Permutation of the columns in any order will generate a new set of peptide sequences, still satisfying the paired exploration condition. The generator described in Eq. 3 generates a compressed peptide library, ensuring that at all possible position pairs of the sequence, all possible binary aminoacid tuples have occurred atleast once. With a minimum redundancy of two for 20 naturally occurring amino-acids, and for 9 length peptide sequence, the size of this peptide library is 1012.

2.6 Implicit structure prediction

The previous section is restricted to $\approx$ 16 000 peptide-HLA complexes whose structures were explicitly modelled. However, in reality, the number of complexes was not restricted to these 16 000, but may run into several billions. Explicit structural modelling is a computationally intensive step. A faster approximation of inter-residue contact patterns is therefore necessary. To address this, a linear optimization problem was framed. For a given peptide-HLA pair, the interaction profile is estimated by choosing weights $\omega_{i,j}$ that maximize the contact potential by using the statistical pair potentials (*CPP*). Bounds on the selection of $\omega_{i,j}$ incorporates structural restriction information.

$$\max_{w_{i,j}} \sum_{i,j} \omega_{i,j} * u(\alpha(p_i), \alpha(h_j))$$

$$\text{Subjected to} : 0 \leq \omega_{i,j} \leq c_{i,j}(\alpha(p_i)) \forall\, i,j$$

$$\sum_i \omega_{i,j} \leq H_j^{\max} \forall\, j$$

$$\sum_j \omega_{i,j} \leq P_i^{\max}(\alpha(p_i)) \forall\, i$$

where

$$u(\bullet, \bullet)$$

represents the *CPP*, as obtained from Eq. 1.

$$c_{i,j}(\alpha(p_i)),\ H_j^{\max}$$

and

$$P_i^{\max}(\alpha(p_i))$$

represent the geometric constraints for each peptide-HLA residue pair interaction, cumulative HLA and peptide residue interactions respectively. An accurate estimate of these geometric constraints was obtained by detailed modelling of each peptide-HLA complex from the *RPL* (Section 2.3).

2.7 Feature representation and binding affinity prediction

The half maximal inhibitory concentration (IC_{50}) is a measure of effective concentration of a substance required to inhibit a particular biological function. The IC_{50} measure for peptide-HLA complex is often used as a yard-stick for evaluating binding strength of an antigenic peptide. At a very low concentration a $\log(IC_{50})$ value closely estimates the Gibbs free energy (ΔG) associated with the binding. The Gibbs free energy associated with the physical binding process has enthalpic and entropic components. These energy contributions arise due to the network of inter-atomic interactions. Thus the change in Gibb's free energy of binding (ΔG) should be a function of all possible atomic interactions at the molecular interface. Implicit structural modelling explained in Section 2.3 yields atomic contact patterns at the peptide interface. These molecular interaction patterns are used to define feature representations. Earlier studies have shown that linear classifiers show poor performance in predicting strong HLA-binding peptides. The exact functional form of the dependencies between different variables are not known. To capture the non-linear nature of the function, Gaussian process regression methods were used.

A Gaussian process defines a 'prior' over functions, which are later converted to a 'posterior' over functions on observation (Rasmussen and Williams, 2006). A Gaussian process for regression resembles the k-neighbourhood predictor, but it provides much more smooth and stable prediction with standard error estimates. Neighbour information is encoded in terms of a kernel function (κ). Kernel functions are positive definite and commutative in nature. A squared exponential kernel, $\kappa(x, x') = \sigma_f^2 \exp(-\frac{d(x,x')}{l^2})$, was used in this study. A Gaussian process assumes that the response variable follows a Gaussian distribution over the observations. For a given set of observations $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ and their reported response value $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, it writes the joint probability distribution

as a Gaussian distribution with zero mean and covariance matrix. Covariances are estimated using a kernel function. Thus for a new observation $(*, *)$, the distribution is given as

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & K_{*,*} \end{bmatrix}\right) \quad (4)$$

This property of Gaussian distribution, enables the estimation of the value for the new observation as $y_*|\mathbf{y} \sim \mathcal{N}(K_* K^{-1}\mathbf{y}, K_{*,*} K^{-1} K_*^{\mathrm{T}})$. The evaluation of the inverse of the covariance matrix is computationally intensive. The Covariance matrix is a positive definite matrix in nature. The computational intensive step of computing the inverse is averted by use of 'Cholesky decomposition'.

This module estimates IC_{50} values for each implicitly modelled complex. The total energy estimate for the system can be approximated as an individual contribution by electrostatic, van der Waals and solvation energy terms. Earlier studies have explored linear models and have found the prediction accuracy to be low (Lafuente and Reche, 2009). Therefore, a non-linear Gaussian process regression scheme is used, with a square exponential kernel.

2.8 Graphical models and representation of epitope pools

The Regression model described in the previous section estimates the binding affinity. The predicted affinity provides a basis for classifying peptide sequences into binder and non-binder classes. This decision is threshold based. 500 nM is considered as a cut-off for strong binder peptides. Regression model is non-linear in nature, and does not provide insight into the mechanistic understanding of the process of binding. A graphical model representation is used to gain an insight to the nature of antigenic peptide sequence profiles.

Graphical models express the positional dependency among peptide residue positions. A probabilistic graphical model (*PGM*) is a tool that helps to represent the conditional dependence structures among random variables, for probabilistic inference for modelling problems involving many variables. *PGM* defines a unique factorization of random variables. In a *PGM*, factorization is a process of presenting joint probabilities for Bayesian inference, in terms of multiplication of conditionally independent joint probabilities of factor graphs as $P(X_1, X_2, \ldots, X_n) = \prod_{i=1\ldots n} P(X_i|\pi(X_i))$, where X_i represents each random variable in the process and $\pi(X_i)$, represents its parent nodes. Each sequence position of the nonameric sequence is modelled as a discrete random variable, which takes a value between $\{1 \ldots 20\}$, corresponding to 20 distinct amino acids. For the appropriate factorization scheme, the dependence structure among the variables must be known. *Chow-Liu* algorithm has been used to extract the dependency structure between the random variables (Chow and Liu, 2006). This algorithm uses mutual information between all pairs of variables. Mutual information (MI) content between two variables describes the extent of dependence between two variables. The variables with strong correlation show a high value for mutual information content, while a zero value represents independence between the variables. *Chow-Liu* algorithm poses the structure determination problem as a two clique problem. It reports a maximum weighted (based on MI content) spanning tree that captures dependencies between different variables. The MI is evaluated from a set of strong binder epitope sequences, as predicted by the non-linear predictor from a pool of a large set of naturally occurring sequences.

The basic assumptions associated with the *Chow-Liu* algorithm are: (i) it cannot handle missing data, (ii) variables are discrete and (iii) data is independently and identically distributed.

3 Results and discussion

3.1 Contact based pair potential (CPP)

The workflow used in this study is illustrated in Figure 1. Since experimental binding affinities are available only for 16 450 peptide-HLA pairs, covering 43 alleles, three dimensional structures of these complexes were modelled (Supplementary Material). For each complex, a contact matrix is computed by counting number of different residue-pairs involved in interactions between peptide (p_i) and HLA (h_j) residues. The frequency of occurrence of such residue-pairs in all complexes are then combined to derive a contact pair-potential(CPP). Figure 2 illustrates the *CPP* matrix, which indicates observed preferences of the 20 amino acids in the peptide (in rows) and in different HLA alleles (in columns). Higher preference for arginine and lysine of the peptides interacting with aspartic acid of the HLA molecule and similarly tyrosine of the peptides interacting with asparagine of the HLA molecule are clearly seen. Figure 2 illustrates the contact potential for all residue-combinations. Conserved residues in the HLA molecules such as the tyrosine can also be seen to interact with several amino acid types in peptides. The derived matrix is asymmetric unlike those for protein-protein interactions and also shows specific residue biases.

3.2 Interaction profiles for each allele

A second line of input that is required for affinity prediction is allele-specific interaction information. 57 different residues form the binding site. Of these it is well known that, at least 21 positions vary significantly across alleles. Therefore the interaction profiles for individual alleles will provide a more accurate picture of the nature and affinity of the peptides it can recognize. Experimental information about the peptide ligands is not available for all alleles. A representative peptide library was first designed. The library was designed with a goal of (i) covering all possible tuples in the peptide size and (ii) covering interactions of all possible residue-pairs between the peptide and the HLA molecules. The representative peptide library (*RPL*) consisting of 1012 peptides was generated (Supplementary Material). For each allele, a complex was modelled with each peptide in the *RPL*, ultimately yielding 2010×1012 complexes. The number of polar and non-polar contacts were calculated in each case and summarized over each allele to output a $57 \times 20 \times 9$ matrix. From this, the contact range statistics for each allele are derived and the maximum values in the range serve as bounds or structural constraints for the next step in the workflow (Fig. 1). The bounds derived for one sample allele each in the **A**, **B** and **C** loci is shown in Figure 3. The figure shows the strength of interaction for each residue-pair between the two molecules, for example alleles. Similar figures can be obtained for all alleles through an interactive query using the web-resource http://proline.biochem.iisc.ernet.in/HLaffy. On an average a binding site residue in the HLA molecule is seen to interact with four peptide residues and ranges from one to six depending on its position. Similarly, the number of HLA residues a peptide residue can interact with is decided by its position on the peptide. The HLA bound peptide conformation is generally seen to have a central bulge, because of which residues **P4** and **P5** have fewer interactions restricting the number of contacts to three, whereas for **P2** and **P9** positions, which serve as anchor residues, interactions with six residues on an average are observed. **P1** and **P8** residues are also located deep inside the grooves, consistent with a higher number of hydrophobic interactions as compared to those in the middle of the peptide.

3.3 Implicit interaction modelling to cater to any peptide-HLA complex

The previous section provides information about allele-specific interaction profiles, computed from explicit structural models of 16

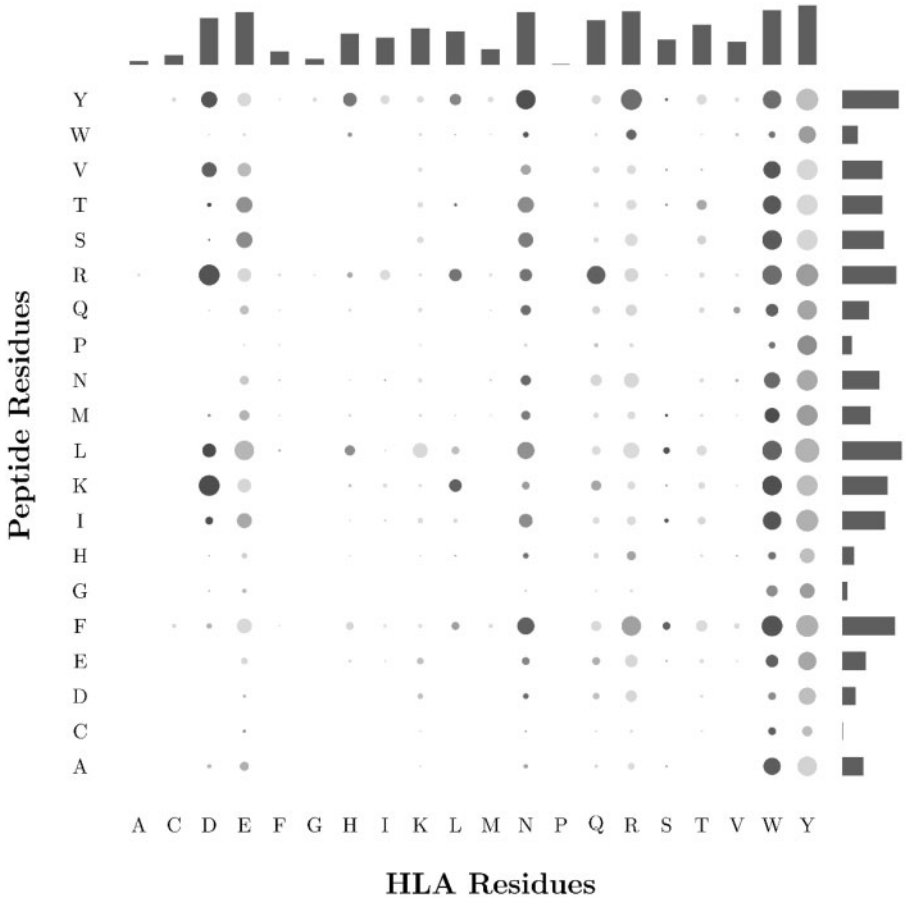

Fig. 2. Contact pairs derived from complexes of protein crystal structures and modelled structures with strong binding peptides. Rows represent peptide residue types. Residues in the HLA binding groove are represented as columns. Circle sizes correspond to the score of the contact pair potential matrix. Cumulative frequencies of amino acids in peptide and HLA binding grooves are shown in the histograms, which are shown along the side and top margin respectively

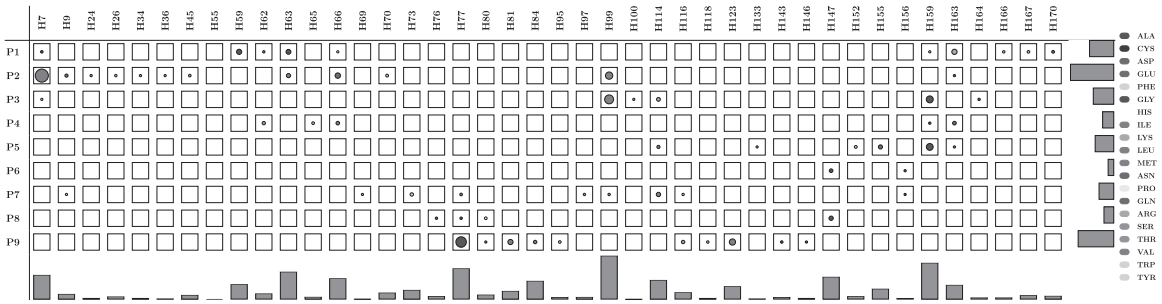

Fig. 3. Geometric constraints obtained from the structure modeling of the RPL. For each HLA sequences, the relative contact between peptide and HLA residues, are shown in each block as a circle. The larger the circle diameter, stronger is the atomic contact as derived from structural modeling. Cumulative interaction count of each peptide residue with HLA molecule is represented in the last column as a histogram. Similarly cumulative residue-wise peptide contacts for the HLA sequence is shown in the histogram at the bottom

000+complexes. Since it is impractical to explicitly model billions of complexes, implicit modelling was carried out for the peptide in the groove of a given HLA molecule. This amounts to implicitly estimating the most feasible position and conformation of the peptide residues in the context of the HLA molecules. This is posed as a linear optimization problem, where the contact potential is maximized for a given peptide-HLA pair, using constraints from the previous module. This exercise yields an approximate estimate of the contacts of the peptide residues in the HLA groove for each peptide-HLA pair, and thereby provides a very large resource of peptide-HLA pairs. To validate how well the implicit modelling fares, we test if interaction counts of complexes with explicit structural models are reproduced well. 16 450 complexes with explicit structural models were taken and the interaction counts for the corresponding pairs from the large peptide-HLA dataset generated in this step are compared. A difference matrix is then generated, which shows similarities greater than 70% in their contact profiles. This exercise therefore indicates that the implicit modelling captures the

interaction profiles of peptide-HLAs reasonably well, with an added advantage of a much larger coverage of the peptide space and caters to any given HLA allele.

3.4 Predicting IC_{50} values for a given peptide-HLA pair

From the previous module, interaction profiles are obtained, which are used to construct the feature vectors for each peptide-HLA complex. The feature vectors represent all possible interaction pairs at each position of the peptide. These features are used in the Gaussian Process for estimating binding affinities. The feature vectors are specific to each complex and are also meaningful in capturing differences between various complexes. The Gaussian process yields a close estimation of the binding strength of a given peptide to a given HLA allele (Fig. 4). To validate the performance of the method, the following analyses were carried out:

i. IC_{50} value predicted by *HLaffy* was compared with an independent dataset, with experimentally derived affinities, and the correlation was found to be 0.85 and a prediction accuracy of 92% (Supplementary Material).
ii. A 5-fold cross-validation exercise was performed for the IEDB dataset with (8 500 complexes), and the average prediction accuracy was found to be 82.5%.
iii. HLaffy performance was compared with three other available methods from literature, which are *ANN*, *NetMHCpan* and *SMM*, for the same dataset. The receiver operator curve (ROC) shown in Figure 4b, which indicates that the performance of ***HLaffy*** is significantly improved as compared to the other three methods (supplementary material). Prediction accuracies for specific alleles were computed for 31 alleles. The best performance was observed for alleles *A*32:01*, *A*02:06* and *B*15:03* are over ≥ 0.85. *B*54:01* allele shows low Pearson correlation of 0.73 (Supplementary Material).

3.5 Graphical models and presentation of epitope pools

A given gene or genome sequence is scanned to identify the set of peptides that can bind to a given HLA allele. This has been made available for web-interactive querying at HLaffy. By default all those peptides that have estimated IC_{50} values of $\leq$ 50 nM are considered as possible epitopes for the given allele. The value is chosen because it is a well-accepted cutoff for strong binders (Roomp *et al.*, 2010). A given sequence can be queried for any of the 1000 alleles. Conversely, any given sequence can be queried for a given allele as well. As an example, the entire set of 4300 viral genomes (1 99 708 sequences) are taken from the NCBI repository and scanned for epitopes of 3 different alleles, one from each of **A**, **B** and **C** loci.

So far, the analyses have captured residue preferences for individual peptide-HLA residue-pairs, in line with an additive model. However, the linear optimization along with the CPPs computed earlier indicates that some positions show a trend of dependent preferences between them. In other words, there can be scenarios where a residue at position 4 can have a strong preference for the residue type at position 7. From the set of peptides predicted as strong binders for each of these alleles individually, probability matrices capturing the preferred residues at each of the nine peptide positions are written out. These are represented in the form of sequence logos and are shown in Figure 5. In this example, a strong preference for *T*, *D* and *Y* residues at positions 2, 3 and 9 respectively are seen for allele **A*01:01**. On the other hand strong preferences for *P*, *R* and *L* at positions 2, 3, 9 are seen for the **B*07:02** allele. Positions 2 and 9 of the peptide are well known to serve as anchors and have strong residue preferences. This analysis identified these as expected, but in addition strong preferences are identified at other positions as well for different alleles. Knowledge of such influences is extremely useful for first gaining a mechanistic understanding of the precise interactions that contribute to binding, which is then directly applicable for predicting peptides for a given allele.

3.6 Estimating epitope pools

The number of peptides predicted from a given genome varies considerably for different alleles. A large scale epitope prediction exercise was carried out. Epitopes were predicted for 1110 distinct alleles, with 359 **A**, 562 **B** and 189 **C** alleles respectively. The analysis shows that on average that more number of epitopes are presented by **A** alleles compared to **B** or **C** alleles. Epitope prediction over all known viral proteins, shows that the number of epitopes presented by **A*02:11**, **A*02:69** and **A*30:04** is significantly higher (≈9000 peptide antigens) compared to other alleles. Among **B** alleles, **B*15** alleles show a high degree of antigen presentation strength, with an antigen pool size as large as 6000 peptides. Among **C** alleles **C*03** loci shows a high degree of peptide recognition. 46 out of the 1110 alleles can recognize as many as 4000 distinct epitopes.

4 Conclusion

In this study, we present a new method HLaffy, for predicting peptide ligands for HLA class-1 molecules, based on a multi-module

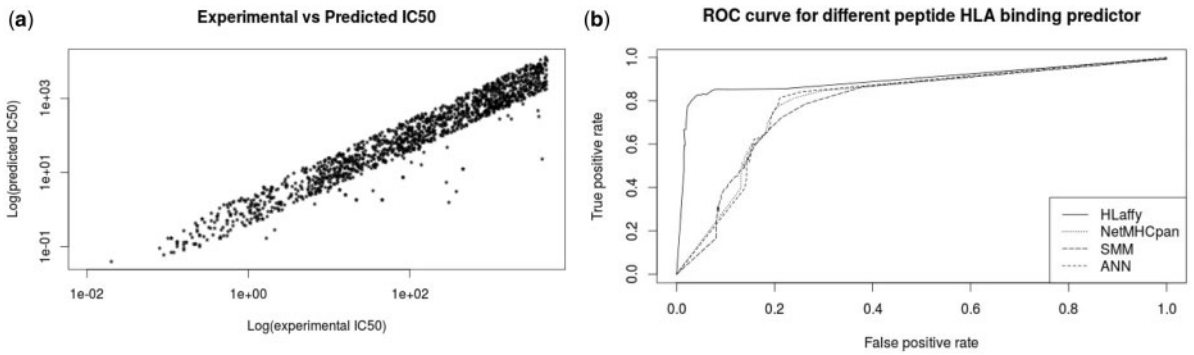

Fig. 4. Benchmarking and validation of HLaffy for accuracy of peptide-HLA binding affinity prediction. (**a**) Predicted IC_{50} using Gaussian Process regression shows strong correlation with experimentally obtained IC_{50} values drawn as a log–log plot. (**b**) Receiver Operator Characteristic curve for comparative study of prediction accuracies among different prediction methods for the same dataset

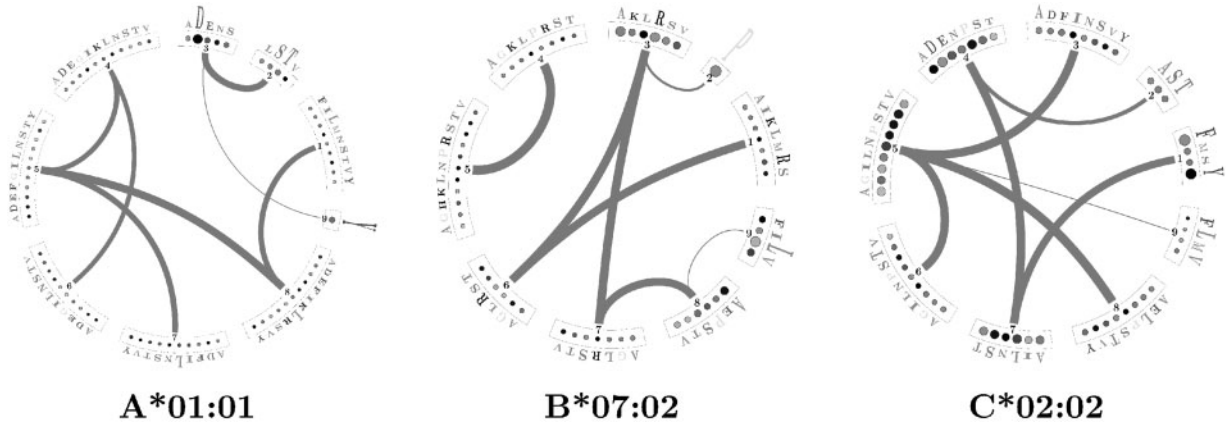

Fig. 5. 2D sequence logo of binder peptides for **A*01:01**, **B*07:02** and **C*02:02** alleles. Circles are shown for each of the nine peptide positions along the circumference, tagged with frequently occurring amino acids, whose sizes are based on the mutual information. Arcs joining two positions indicate an influence between them

workflow. HLaffy first obtains a mechanistic understanding of the recognition specificity and is capable of explaining experimentally observed binding affinities for a large dataset of peptides. It then identifies the highest contributing pair potentials and learns cross-residue influences through graphical models. The implicit structural modelling is a new contribution towards understanding peptide recognition by HLA molecules and estimating their binding strengths. This leads to a phenomenal increase in coverage of the peptide space. Another novel aspect of this study is to obtain a high correlation of observed IC_{50} values with interactions, which has been possible because of weighting different interactions and estimating their relative contributions through feature representation. A high correlation implies that the IC_{50} values can be rationalized from the structures of peptide-HLA complexes. Sequence logos used widely in literature as a representation of the sequence patterns of epitopes is an elegant method to identify individual residue preferences but are not directly useful for identifying inter-residue dependencies. Besides these, another contribution from the study is a novel design strategy for constructing a representative peptide library. The graphical model used in this study is capable of capturing such dependencies systematically. HLaffy with its ability to sample the peptide space more comprehensively opens up new possibilities to explore the repertoire of epitopes that can be recognized by a given allele, providing estimates of the total epitope pool sizes for each allele. The importance of predicting epitopes for HLA molecules is well recognized in the field since it can be applied in diverse areas such as vaccine discovery, understanding disease susceptibilities, autoimmune disorders and in estimating success of organ transplants.

Author contributions

SM developed and implemented the prediction algorithm with guidance from NC and CB through the study. SM and NC wrote the paper and all authors read and approve the final manuscript.

Funding

We gratefully acknowledge a financial grant from Department of Science & Technology, Government of India (DST), for the Centre of Excellence Mathematical biology (DSTO/PAM/GR-1303).

Conflict of Interest: none declared.

References

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhasin,M. *et al.* (2003) Mhcbn: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 665–666.

Chow,C. and Liu,C. (2006) Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theor.*, **14**, 462–467.

Doytchinova,I.A. *et al.* (2004) Identifiying human MHC supertypes using bioinformatic methods. *J. Immunol.*, **172**, 4314–4323.

Hess,B. *et al.* (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

Hewitt,E.W. (2003) The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, **110**, 163–169.

Hobohm,U. and Meyerhans,A. (1993) A pattern search method for putative anchor residues in T cell epitopes. *Eur. J. Immunol.*, **23**, 1271–1276.

Hoof,I. *et al.* (2009) Netmhcpan: a method for MHC class I binding prediction beyond humans. *Immunogenetics*, **61**, 1–13.

Krivov,G.G. *et al.* (2009) Improved prediction of protein side-chain conformations with scwrl4. *Proteins*, **77**, 778–795.

Kumar,N., and Mohanty,D. (2007) Modpropep: a program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res.*, **35**, W549–W555.

Lafuente,E.M., and Reche,P.A. (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. *Curr. Pharm. Des.*, **15**, 3209–3220.

Lundegaard,C. *et al.* (2008) Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, **24**, 1397–1398.

Lundegaard,C. *et al.* (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology*, **130**, 309–318.

Murphy,K.M. *et al.* (2007). *Janeway's Immunobiology (Immunobiology: The Immune System (Janeway))*, 7th ed. Garland Science, New York.

Parker,K.C. *et al.* (1994) Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.

Peters,B. and Sette,A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, **6**, 132.

Rammensee,H. *et al.* (1999) Syfpeithi: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.

Rasmussen,C. and Williams,C. (2006). *Gaussian Processes for Machine Learning. Adaptative Computation and Machine Learning Series*. University Press Group Limited, The MIT Press, Cambridge.

Robinson,J. *et al.* (2013) The imgt/hla database. *Nucleic Acids Res.*, **41**, D1222–D1227.

Roomp,K. *et al.* (2010) Predicting MHC class I epitopes in large datasets. *BMC Bioinformatics*, **11**, 90.

Sali,A., and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Schueler-Furman,O. *et al.* (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.*, **9**, 1838–1846.

Sette,A. and Sidney,J. (1999) Nine major HLA class I supertypes account for the vast preponderance of hla-a and -b polymorphism. *Immunogenetics*, **50**, 201–212.

Van Der Spoel,D. *et al.* (2005) Gromacs: fast, flexible, and free. *J. Comput. Chem.*, **26**, 1701–1718.

Vita,R. *et al.* (2010) The immune epitope database 2.0. *Nucleic Acids Res.*, **38**, D854–D862.

Yanover,C. and Bradley,P. (2011) Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 6981–6986.