Nagpal *et al. J Transl Med (2018) 16:181*
https://doi.org/10.1186/s12967-018-1560-1

Journal of
Translational Medicine

RESEARCH

Open Access

Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants

Gandharva Nagpal[1†], Kumardeep Chaudhary[1†], Piyush Agrawal[1] and Gajendra P. S. Raghava[1,2*]

Abstract

Background: Evidences in literature strongly advocate the potential of immunomodulatory peptides for use as vaccine adjuvants. All the mechanisms of vaccine adjuvants ensuing immunostimulatory effects directly or indirectly stimulate antigen presenting cells (APCs). While numerous methods have been developed in the past for predicting B cell and T-cell epitopes; no method is available for predicting the peptides that can modulate the APCs.

Methods: We named the peptides that can activate APCs as A-cell epitopes and developed methods for their prediction in this study. A dataset of experimentally validated A-cell epitopes was collected and compiled from various resources. To predict A-cell epitopes, we developed support vector machine-based machine learning models using different sequence-based features.

Results: A hybrid model developed on a combination of sequence-based features (dipeptide composition and motif occurrence), achieved the highest accuracy of 95.71% with Matthews correlation coefficient (MCC) value of 0.91 on the training dataset. We also evaluated the hybrid models on an independent dataset and achieved a comparable accuracy of 95.00% with MCC 0.90.

Conclusion: The models developed in this study were implemented in a web-based platform VaxinPAD to predict and design immunomodulatory peptides or A-cell epitopes. This web server available at http://webs.iiitd.edu.in/raghava/vaxinpad/ will facilitate researchers in designing peptide-based vaccine adjuvants.

Keywords: Immunomodulatory peptide, Antigen presenting cells, A-cell epitopes, Support vector machine, Adjuvants

Background

Peptide subunit vaccines are hailed as an advancement over live or inactivated whole organism vaccines due to their ability to minimize adverse reactions [1]. Yet, antigenic peptides by themselves are poorly immunogenic since they lack the capability of activating the innate immunity. Activation of the innate immune system is required for stimulation of whole immune system including adaptive immunity. Hence, there is a need for inclusion of immunostimulants known as adjuvants in the subunit vaccine formulations. Conventionally, empirical approaches were used for adjuvant discovery, so far limited adjuvants have been approved and licensed for clinical use like alum, MF59, AS03 and AS04 [2].

Vaccine adjuvants effectuate their action by a variety of mechanisms with all of them involving the antigen presenting cells (APCs) particularly the dendritic cells [3]. One of these mechanisms is the activation of the pattern recognition receptors (PRRs) on the APCs that recognize conserved microbial molecular signatures. PRR ligands shape the adaptive immune response mediated by the APCs. A majority of the vaccine adjuvants are ligands of PRRs making them potential targets for rational design of vaccine adjuvants [2]. Thus, hypothesis-driven adjuvant

*Correspondence: raghava@iiitd.ac.in
†Gandharva Nagpal and Kumardeep Chaudhary contributed equally to this work
[2] Centre for Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India
Full list of author information is available at the end of the article

development relies on the expectation that the mechanistic understanding of the immune responses exhibited by PRR ligands would enable fine-tuning the specificity of adjuvants to attain vaccine efficacy and safety, simultaneously. An important example of a class of molecules that have been shown to have immunomodulatory effects and are poised to become safe and cost-effective adjuvants in future is—short immunomodulating peptides [4]. Figure 1 is a schematic representation of the adaptive immune cell activation by a coordination of antigen presentation to the naïve adaptive immune cell with the release of cytokine milieu mediated by PRR activation. Keeping in view the role of peptide ligands of PRRs in the activation of APCs, we introduce the term 'A-cell epitopes' for these immunomodulatory peptides.

Cationic host defense peptides (HDPs) were originally discovered as antimicrobial peptides produced within the multicellular organisms having a broad-spectrum activity against bacteria, viruses, fungi, protozoa, etc. [5]. Of late, HDPs and their synthetic analogs called innate

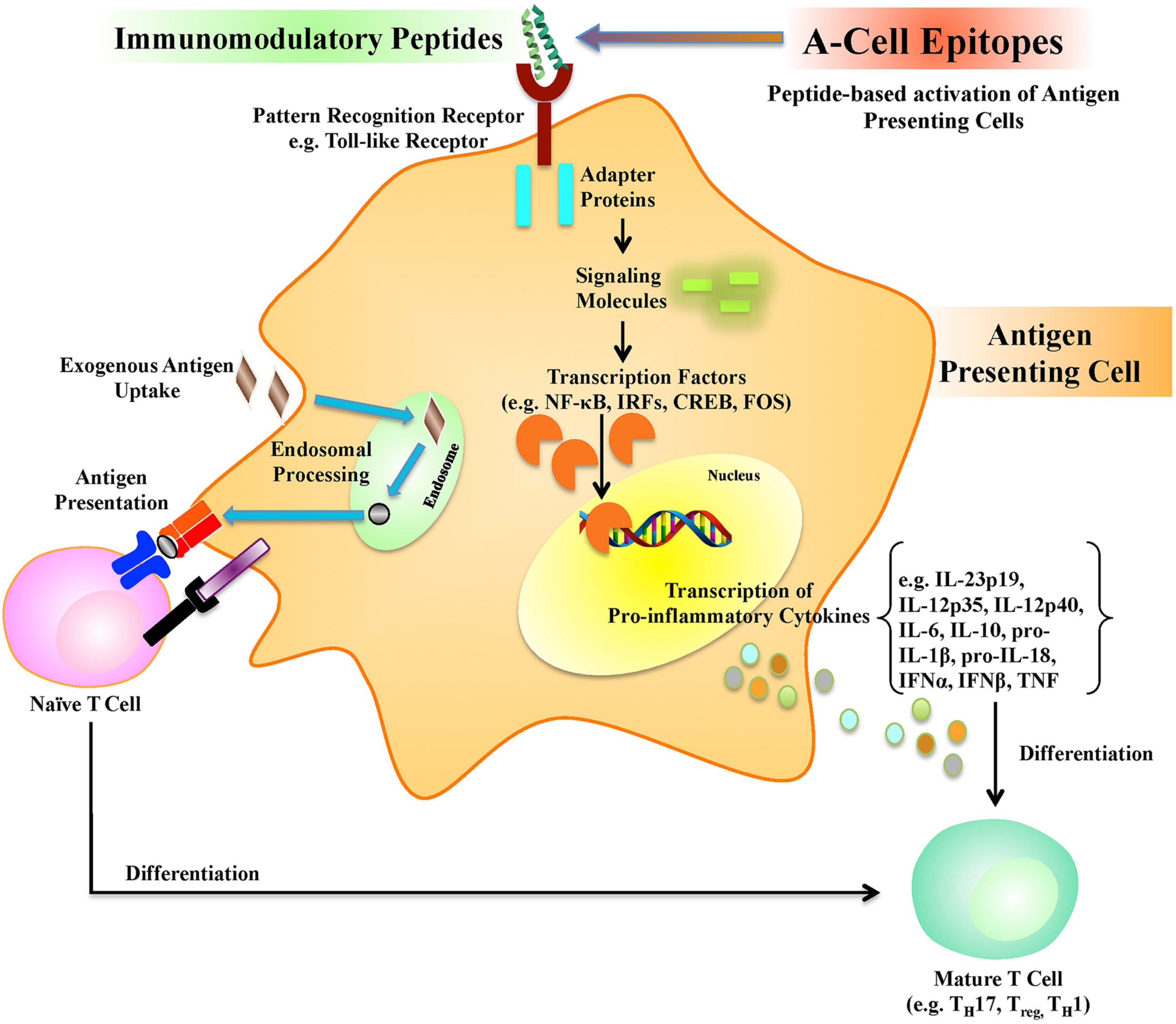
Immunomodulatory Peptides
A-Cell Epitopes
Peptide-based activation of Antigen Presenting Cells
Pattern Recognition Receptor e.g. Toll-like Receptor
Adapter Proteins
Signaling Molecules
Transcription Factors (e.g. NF-κB, IRFs, CREB, FOS)
Antigen Presenting Cell
Exogenous Antigen Uptake
Endosomal Processing
Endosome
Antigen Presentation
Nucleus
Transcription of Pro-inflammatory Cytokines
e.g. IL-23p19, IL-12p35, IL-12p40, IL-6, IL-10, pro-IL-1β, pro-IL-18, IFNα, IFNβ, TNF
Differentiation
Naïve T Cell
Differentiation
Mature T Cell (e.g. T_H17, T_{reg}, T_H1)

Fig. 1 An illustrative mechanism of antigen presenting cell (APC) activation caused by immunomodulatory peptides through innate immune receptors leading to the induction of adaptive immune cells. The immunomodulatory peptides are ligands of innate immune receptors that evoke cytokine expression through cellular signaling pathways. The cytokines lead to the maturation of naïve cells into mature adaptive immune cells such as various types of T-lymphocytes. Since the immunomodulatory peptides activate the APCs leading to the activation of the adaptive immune cells, they may be used as vaccine adjuvants and be called 'A-cell epitopes'. The figure was drawn using ScienceSlides, made available at http://www.scienceslides.com/ by VisiScience

defense regulators (IDRs) have been realized to cause immunomodulatory effects like differentiation and activation of innate and adaptive immune cells, modulation of pro- and anti-inflammatory responses, chemo attraction, autophagy, apoptosis and enhancement of immune-mediated bacterial killing [6]. Many host defense peptides (HDPs) with known immunomodulatory effects are already in clinical trials [7]. IMMACCEL-R is a short synthetic peptide with immunomodulatory properties that has been commercialized for use as vaccine adjuvant in animals and birds for the purpose of antibody generation [8]. The human cathelicidin antimicrobial peptide (CAMP or LL37) is another well-known antimicrobial peptide shown to induce immunomodulatory effects [9] and has been found to be associated with immune-related disorders like psoriasis [10] and morbus Kostmann [11].

Adjuvants have been incorporated into the vaccine formulations for qualitative alteration of the adaptive immune responses that are different from non-adjuvanted antigens. For instance, the adjuvants have been used to skew the immune responses with respect to Th1 (T-helper 1) cells versus Th2 cells, $CD8^+$ versus $CD4^+$ T cells, specific antibody types, etc. [2]. The PRR ligands produce these effects by virtue of the adapter proteins in the signaling pathways activated by the PRR. For example, bacterial flagellin protein causes adjuvant effect through TLR 5 and produces a mixed Th1 and Th2 response instead of polarized Th1 response and requires TLR adapter protein MyD88 for this effect. In contrast, monophosphoryl lipid A (MPL) and bacterial lipopolysaccharide (LPS) acting through TLR 4 activation lead to production of pro-inflammatory cytokine TNF leading to a polarized Th1 response instead of mixed Th1–Th2 response. While MPL signals through TRIF adaptor, LPS mediated activation of TLR 4 acts through both TRIF and MyD88 adapter proteins. Thus, MPL formulated on alum (AS04) stimulates a polarized Th1 cell response and is a component of licensed vaccine for HBV and papilloma that has proven to be both safe and effective.

Developing adjuvants based merely on empirical studies without the understanding of mechanisms is inadequate [2]. There is a need to develop systematic and rational approaches for designing highly potent vaccine adjuvants. One such approach could be the development of PRR ligands into vaccine adjuvants since their mechanism is known.

In such a scenario, in silico models to screen and identify potential vaccine adjuvant candidates could prove to be useful as the existing experimental approaches are time and resources consuming [12]. Previously, our group developed a method, Vaccine DA for predicting immunomodulatory oligodeoxynucleotides that can activate innate immune system via Toll-like receptor-9 (TLR-9). This tool can be used for designing oligodeoxynucleotide-based vaccine adjuvants as well as for genome-wide screening of vaccine adjuvants [13]. Recently, we also developed a method imRNA for designing single-stranded RNA (ssRNA) based vaccine adjuvants [14]. These methods may play an important role in designing DNA and RNA-based therapeutics as these methods allow a user to design oligonucleotides of desired immunogenicity.

In the last two decades, numerous method have been developed for predicting potential of peptides to stimulate adaptive arm of the immune system that include methods for predicting MHC binders, B-cell epitopes [15–23] and T-cell epitopes [24–31]. To the best of our knowledge, no method has been developed so far for predicting immunostimulatory potential of peptides to activate innate immunity. In this study, we made an effort to develop method for predicting immunomodulatory peptides that can activate innate arm of immune system or antigen presenting cells. These peptides activate the antigen presenting cells (e.g., dendritic, macrophages); hence, we propose that these immunomodulatory peptides be termed as 'A-cell epitopes'.

In the present work, first we collected experimentally identified immunomodulatory peptides from the literature and included them in our positive set named A-cell epitopes. Next, we collected the human endogenously circulating peptides to build the negative set named A-cell non-epitopes. Combining the positive and the negative sets into a complete dataset, we developed support vector machine (SVM) based computational models that can classify a new query peptide as A-cell epitope or non-epitope. To benefit the users of the scientific community, we provided the best performing SVM-based prediction models in the form of a web-based application called VaxinPAD to be used for identifying and designing novel A-cell epitopes. Such peptides identified computationally might serve as the starting molecules for designing peptide-based vaccine adjuvants.

Methods

Dataset

The experimentally validated immunomodulatory peptide sequences were obtained from 16 patents. As an example of the sequences considered immunomodulatory, a set of sequences taken from a patent (US20110008318 A1), includes flagellin-derived peptides that exhibit immunomodulatory effect by direct binding to TLR 5 as indicated by assays reporting increased NF-κB expression estimated from coupled luciferase activity and TNFα production estimated using flow cytometry. In one of the patents (US7462360 B2), a class of immunomodulatory peptides, called alloferons, derived from the bacteria challenged blood of larvae of the insect blowfly, *Calliphora vicina*, have been found to stimulate the cytotoxic anticancer activity of the human NK-cells and lymphocytes. In another case, a set of peptides as described in patent US8791061 B2 have been shown to enhance innate immunity by modulating the activity of type II transmembrane serine protease dipeptidyl peptidase (DPPIV) also known as CD26 or adenosine deaminase binding protein, expressed on major immune cells like activated T-cells, B-cells, NK-cells, macrophages and epithelial cells. With two major functions of signal transduction and proteolysis, the effects of DPPIV protein-mediated cellular processes include modulation of the chemokine activity by cleaving dipeptides from chemokine N-terminus that alters the receptor binding and specificity of the processed chemokine. DPPIV is a neutrophil chemorepellant and eosinophil chemoattractant too.

After removing the longer sequences, 304 unique sequences left in the length range of 3–30 residues were used to constitute the positive dataset named here as the A-cell epitopes. The upper bound of length 30 residues was kept as more than 90% of the originally collected epitope sequences were retained keeping this criterion used for removing very long sequences. In the absence of experimentally verified non-immunomodulatory peptides (non-epitopes), the experimentally identified endogenous human serum peptides [32, 33] were taken as non-epitopes. We assume these peptides are non-immunogenic as they are part of human serum, thus we assign them as non-epitopes. Only the sequences of the length 3–30 were taken into the negative dataset. In this manner, the main dataset consisted of 304 A-cell epitopes and 385 non-epitopes. Additional file 1: Table S1 provides the sequences and the source patent/publication for the positive and the negative datasets.

Input features

In order to develop any in silico model it is important to generate input features corresponding to each data point. In this study, a data point is the amino acid sequence of a peptide (either A-cell epitope or non-epitope). It is important to generate fixed length input features because machine-learning techniques require fixed length vector for developing a model. As the length of peptides is variable, thus we computed amino acid composition of A-cell epitopes and non-epitopes for developing models. We also computed the average amino acid composition of A-cell epitopes, non-epitopes and the human proteins, in order to understand compositional bias in A-cell epitopes. The amino acid composition for each sequence constituted the input vector of length 20, which was used for developing SVM-based prediction models. Similarly, the dipeptide composition vectors of length 400 were generated for A-cell epitopes and non-epitopes with each element of a vector corresponding to the composition value of each type of possible dipeptide. In addition to compositional features, we also generated binary features for developing models using fixed length of amino acids from the termini (N-terminal or C-terminal or both) of peptides. In the case of binary feature, an amino acid is represented by a vector of 20, where the presence of amino acid is indicated by '1' and the absence is presented by '0' [34]. This means a peptide of length N is presented by a vector of length N × 20 in the case of binary features.

Motif search

We used the Motif—EmeRging and with Classes—Identification (MERCI) Program [35] to identify motifs exclusively occurring in the A-cell epitopes [36]. Though this program allows searching for gapped and ungapped motifs, but we restricted our analysis to the ungapped motifs. It is well established that in the case of T-cell epitopes, even a single residue mutation changes its immunogenicity [37] and can even eliminate the immunogenicity of the epitope [38]. Hence, intuitively the ungapped motifs found to be conserved among the positive sequences are more likely to help identify novel A-cell epitopes. Thus, we computed and compared the frequency of occurrence of ungapped motifs in A-cell epitopes, non-epitopes and the Swiss-Prot proteins.

Classifiers based on machine learning techniques

In the present study, some commonly used popular machine learning techniques were used to develop classification prediction models. We used WEKA package to implement these machine learning techniques namely Random Forest, Naïve Bayes, SMO and J48 [39]. These classification models were developed using commonly used features of peptides like amino acid composition (AAC) and the dipeptide composition (DPC).

Support vector machine (SVM)

Subsequent prediction models in this study were developed using SVM, which has been frequently used to develop models for epitope prediction in previous studies [21, 23, 28]. SVM has been the method of choice for building epitope prediction models especially T-cell epitopes [40] due to its ability to provide effective models on high dimensionality data with less data points. Also, in the past studies it has been shown that SVM performs better on independent dataset in comparison to other machine learning classifiers [41]. The dataset used in the current study contains data points comparable in number to the dimensionality. Hence, we optimized the prediction models on various parameters using the radial basis kernel of a freely available program SVM^{light} [42] to select the best performing models on different sets of features.

Evaluation of models using internal and external validation

In this study, standard procedure was followed to evaluate the performance of models in order to avoid biases in performance due to over optimization. Our main dataset was divided into two categories internal and external dataset, where the internal dataset contained ~80% sequences and the external dataset comprised of the remaining 20% sequences. In order to perform internal validation, we performed fivefold cross validation technique on internal dataset. In this technique, the dataset is divided in five sets, four sets are used for training a model, and the remaining set is used for testing the model. This process is repeated five times so each sequence is tested only one time. In order to perform the external validation of a model, the best model developed using fivefold cross validation is tested on an external dataset. It is important to assess the performance of a model on external or independent dataset because the performance of a model in internal validation may be biased due to optimization of the model [28]. The performance of models was measured using standard threshold dependent parameters namely sensitivity, specificity, accuracy and Matthew's correlation coefficient (MCC) [19, 36] and a threshold independent parameter area under receiver operating characteristics (AUROC) [43].

Bootstrap aggregating

In order to avoid over fitting of models and reducing variance in performance of models; we used bootstrap aggregating (bagging) for averaging performance of models. In this study, process of creating internal and the external datasets has been repeated ten times. Each time, the sequences for the internal dataset were randomly selected from the main dataset, and the remaining sequences were included in external dataset. Finally, we evaluated the performance of our models using various features on both the internal as well as the external datasets as described in above sections. This process gave 10 performance values using internal and 10 performance values using external validation from 10 rounds of sampling. We computed the mean and standard deviations of these performance values to check for bias in performance of the models depending on the choice of sequences on which the models were trained or independently evaluated.

Random peptides as negative dataset

As described above, initially the negative dataset consisted of the experimentally identified endogenous human serum peptides as non-epitopes constituting the negative dataset. We further wanted to check whether the performances of the classification models were dependent on the choice and size of the negative datasets. This was necessary as the negative dataset does not contain the experimentally verified non-epitopes. For this, we created an alternative negative dataset of random peptides derived from the human proteins obtained from the Swiss-Prot database. As mentioned in the previous section, for each of the 10 rounds of sampling, a different set of random peptides 10 times the number of the positive sequences (A-cell epitopes) from the human proteins was kept as the negative dataset.

Results

Compositional analysis

One of the objectives of this study is to understand the nature of A-cell epitopes regarding the residues preferred in A-cell epitopes. Thus, we computed the average residue composition of A-cell epitopes and the non-epitopes. The non-epitope dataset consists of peptides occurring in the normal human serum assumed to be non-immunomodulatory. In addition, the average residue composition of the Swiss-Prot Human proteins was also computed and compared with that of the A-cell epitopes.

In the A-cell epitope dataset, the percentage composition of an amino acid residue was calculated for each epitope, and the average of these values was plotted in Fig. 2 for the corresponding amino acid. Similarly, the average percentage composition was calculated for all the amino acids in the non-epitope dataset and the Swiss-Prot Human proteins. As shown in Fig. 2, the residues showing noticeable differences in average composition between A-cell epitopes and non-epitopes are C, D, E, I, L, R, S, T, V and W. Student's t-test significance value (p-value) was calculated for each residue type to check whether the composition values among A-cell epitopes were different from those in the non-epitopes. In decreasing order of significance (increasing adjusted p-value), the residues R, E, T, S, D, V, W, L, I and C showed the most significant difference between the A-cell epitopes and non-epitopes among all of the residue types with adjusted p-values 1.76E−39, 1.66E−24, 7.04E−18, 3.59E−17, 3.72E−12, 2.79E−11, 5.21E−10, 7.91E−10, 1.24E−09, 1.76E−08 respectively (Additional file 1: Table S2). In particular, when compared to the human proteins taken from Swiss-Prot; R was found to have a higher average composition in A-cell epitopes. The average composition of R in non-epitopes is lower than

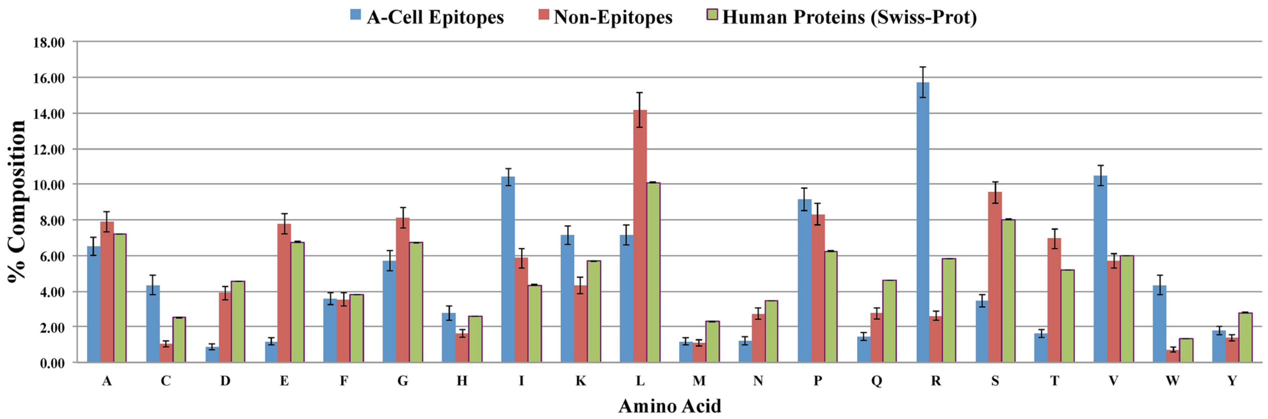

Fig. 2 Barplots showing the comparison of percent average amino acid composition of A-cell epitopes (blue) with non-epitopes (red) and Swiss-Prot human proteins (green)

that in the human proteins. Overall, the residues I, R, V and W were found to be more abundant in the A-cell epitopes as compared to the non-epitopes and Swiss-Prot Human proteins.

Similarly, the dipeptide and tripeptide compositions of the A-cell epitopes and non-epitopes were also compared with the Swiss-Prot Human proteins. Additional file 1: Table S3 gives the average composition for each dipeptide in the A-cell epitopes, non-epitopes as well as the Swiss-Prot Human proteins. After sorting the table according to descending order of difference of dipeptide composition between the A-cell epitopes and the human proteins, top 10 dinucleotide include the residues I, R and V. But these motifs also contain other amino acids that show less significant difference of abundance as compared to the non-epitopes and human proteins. Similar analysis of tripeptide composition is shown in Additional file 1: Table S4. In this case too, the top 10 tripeptide motifs include less abundant residues apart from I, R and V.

Terminal residue preference

We performed position-specific analysis of residues in A-cell epitopes to understand the type of residues preferred at different positions in A-cell epitopes. In this study, two-sample logo (TSL) tool (available at http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi) [44] was used to visualize residues preferred or not preferred in A-cell epitopes. Since the minimum peptide length in the dataset was 3, the N-terminal 3 residues of both the negative and the positive sequences were taken as input to build the N-terminus TSL. C-terminus TSL was obtained using the C-terminal 3 residues from the dataset. Figure 3 shows that the residues R, V and I are among the preferred residues in the A-cell epitopes at both the N and the C termini.

MERCI motif analysis

The Motif—EmeRging and with Classes—Identification (MERCI) Program is a software that helps in finding the motifs exclusive to one class when compared to another class of sequences. Additional file 1: Table S5 provides the MERCI motifs exclusive to the A-cell epitopes as compared to non-epitopes. Top 10 ungapped motifs with respect to the occurrence in the A-cell epitope sequences have a frequent occurrence of I, R and V. On the other hand, ungapped MERCI motifs exclusive in non-epitopes (Additional file 1: Table S6) that are top 10 in abundance contain E, G, P and L.

Rare motif occurrence

We compared the occurrence of peptide *n*-mers ($n=3$, 4, 5, 6) in the A-cell epitopes and non-epitopes. First, the occurrence of each type of *n*-mer was counted in all of the Swiss-Prot proteins, and the *n*-mers were arranged in increasing order of occurrence. In this order, the *n*-mers were divided into 8 bins such that the 1st bin contained the *n*-mers least abundant in Swiss-Prot while the 8th bin contained the most abundant *n*-mers occurring in the Swiss-Prot. Next, the percentage of *n*-mers in a particular bin that occur in the Swiss-Prot was calculated with respect to the total number of *n*-mers in Swiss-Prot. Similar percentage value was calculated for A-cell epitopes and non-epitopes for each bin, and the values were presented in the form of a plot in Fig. 4.

Figure 4a shows that tripeptides in the first three bins occur more in A-cell epitopes while those of 4th, 5th, 6th, 7th and 8th bin (most abundant Swiss-Prot tripeptides) occur more in the non-epitopes. For tetrapeptides (Fig. 4b), the bins having more number of tetrapeptides occurring in the A-cell epitopes than non-epitopes are

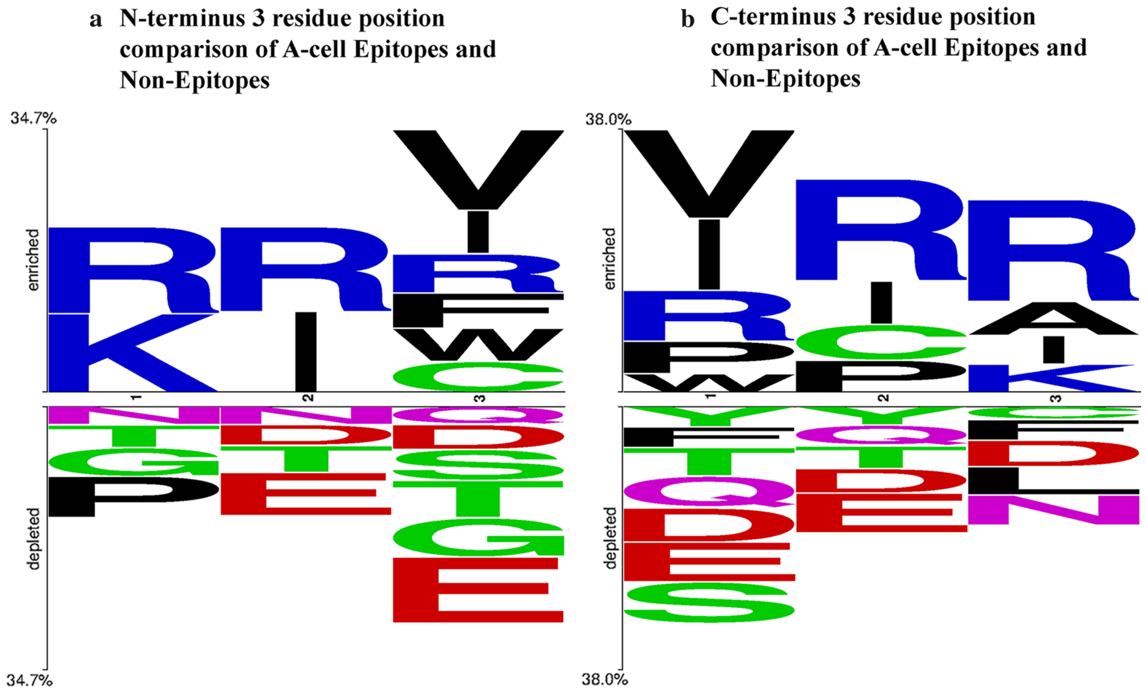

Fig. 3 Two-sample logo of the 3 residue positions at the **a** N-terminus and **b** C-terminus of the A-cell epitopes and non-epitopes. Enriched label represents the positive dataset whereas depleted label represents the negative dataset. In a two-sample logo, the height of a symbol at a residue position is proportional to the difference in symbol frequency between the positive and the negative datasets at that residue position. In the case of A-cell epitopes (as positives) and non-epitopes (as negatives), R is a preferred amino acid at terminal positions apart from I and V. The symbol colors are in accordance with the WebLogo default color scheme provided by the server available at http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi. In the default WebLogo color scheme, residues G, S, T, Y and C appear in green color, N and Q are colored purple, K, R and H are depicted in blue, D and E are drawn red and P, A, W, F, L, I, M and V are shaded black

1st, 2nd, 3rd and 4th. Figure 4c shows the occurrence of the pentapeptides. The bins having distinctly more pentapeptides in the A-cell epitopes than non-epitopes are again the first four bins. On the other hand, the percentage occurrence of hexapeptides of A-cell epitopes is lower than non-epitopes and Swiss-Prot proteins only in the 8th bin (Fig. 4d).

Prediction of immunomodulatory peptides

The sequence-based analyses like residue composition preferences; position-wise residue preference and motif search indicated that these features could help in discriminating the A-cell epitopes from non-epitopes. We developed SVM-based prediction models using SVM*^{light}* by from the dataset of 304 A-cell epitopes as positive sequences and 385 non-epitopes as negative sequences. From each of the positive and negative datasets, ~80% sequences were kept in the training–testing dataset while the remaining ~20% were kept in the independent dataset. Thus, the training–testing dataset had 243 positive and 308 negative sequences. The best performing models were selected on the basis of highest Matthews correlation coefficient values and a minimal difference between the sensitivity and specificity values.

Prediction models based on machine learning techniques

In order to understand, which machine learning technique will be most efficient for predicting A-cell epitopes, models were developed using different machine learning techniques. Initially models were developed using SVM implemented with SVM^{light} and four commonly used techniques (Random Forest, Naïve Bayes, SMO and J48) implemented using WEKA package. These models were developed using amino acid composition (AAC) and dipeptide composition (DPC) of peptide sequences (epitope and non-epitope). As shown in Additional file 1: Table S7, SVM based model performed better than models developed using any other machine learning technique. SVM based models on training dataset obtained MCC values 0.90 and 0.91 for AAC and DPC respectively. Similarly, performance was evaluated on the independent dataset. Thus, in this study, we used SVM for developing models using various features of peptides. The performance of SVM models developed using different features have been shown in Additional file 1:

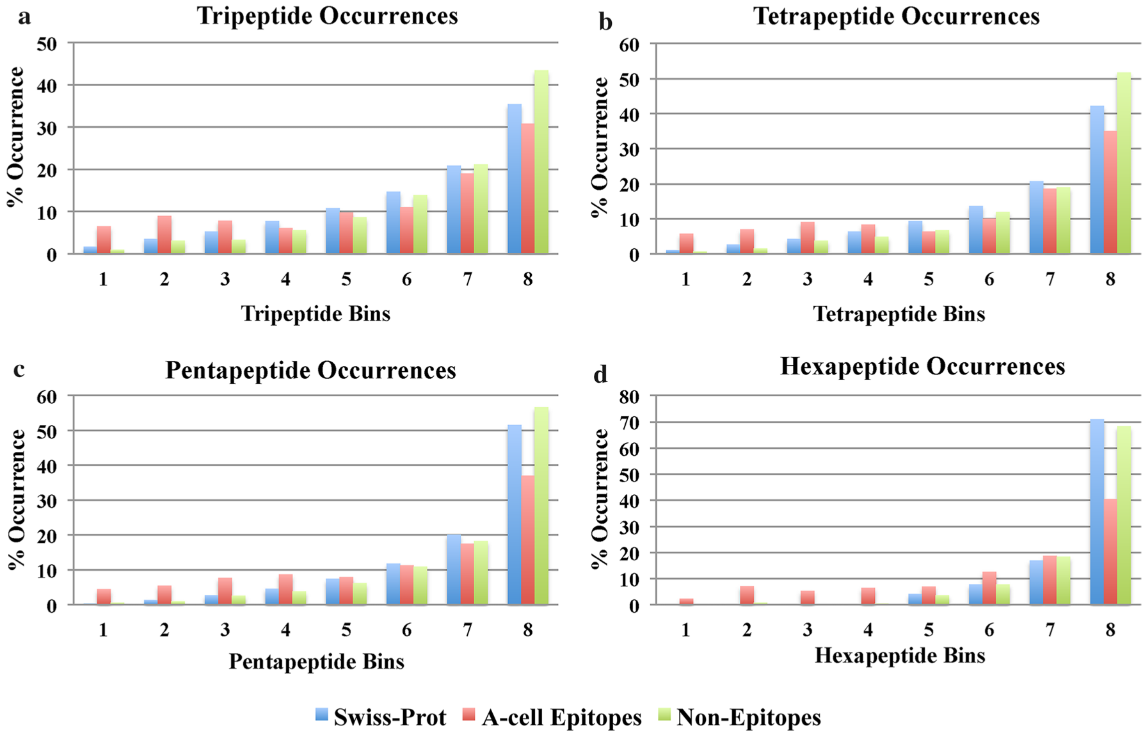

Fig. 4 Comparison of occurrence of **a** tripeptides, **b** tetrapeptides, **c** pentapeptides and **d** hexapeptides divided into 8 bins in the ascending order of occurrence (most rarely occurring to most abundant) in Swiss-Prot proteins

Table S8. A variation was observed in performances of models on some feature sets between the training and the independent datasets. Ideally models performing better on the training dataset should also perform better on independent dataset. In our case, we observed that models performing better on training dataset were performing lower on independent dataset and vice versa (see Additional file 1: Table S8). This inconsistency in models' performance might have arisen from the over fitting of models to the training dataset.

Diminishing over fitting using bagging

In this study, we used bagging approach for sampling to overcome problem of over optimization or over fitting. Bagging procedure of sampling was adopted in this study to evaluate performance of models. The main dataset was divided ten times randomly into the internal and external datasets. Thus, we get ten training/internal datasets and ten independent datasets. The performance of all models was evaluated ten times on training and independent dataset. Finally, we computed the performance of each model in 10 rounds of sampling and reported the mean and standard deviation on these 10 datasets. Table 1 provides the performance values of models developed on various feature sets for the 10 internal dataset samples in different categories.

Composition-based models

The amino acid composition (AAC) and dipeptide (DPC) composition were used to develop SVM-based models on the training–testing dataset. On evaluating the performance parameters, the AAC model gave an accuracy of 93.30% and the Matthews correlation coefficient (MCC) value of 0.87 as given in Table 1. The DPC model also gave a similar performance in terms of accuracy (94.84%) and MCC (0.90) values. When the AAC and DPC models on the terminal 5 residues of the sequences individually (N or C terminus—N5, C5) or together (N and C termini combined—N5C5) were developed, the N5C5 AAC and N5C5 DPC models performed closest to but not better than the AAC and DPC models with MCC values for N5C5 AAC being 0.87 and N5C5 DPC 0.86 (Table 1).

Binary models

Binary models take the residue position into account by representing each residue type as a binary vector. We considered 5 and 10 residue positions from either end of the peptide sequences (N and C termini) and developed SVM-based models individually and in combination. In case of 5-residue position consideration, the model developed on combined 5 residue positions on both the N and C termini (N5C5 bin in Table 1) performed the best giving an accuracy value 90.22% and MCC value

Table 1 The performance of SVM-based models developed using various features; models were evaluated on training dataset using fivefold cross-validation (internal cross-validation)

Features	Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	AUROC	Parameters
AAC	−0.1	94.49 ± 0.80	92.38 ± 1.33	93.30 ± 0.84	0.87 ± 0.01	0.98 ± 0.00	g = 0.001, c = 3, j = 3
N5 AAC	0	88.54 ± 0.75	90.25 ± 1.87	89.44 ± 1.26	0.79 ± 0.02	0.94 ± 0.00	g = 0.0005, c = 2, j = 1
C5 AAC	0	91.13 ± 1.42	92.94 ± 1.20	92.08 ± 1.05	0.84 ± 0.02	0.97 ± 0.00	g = 0.001, c = 9, j = 1
N5C5 AAC	−0.2	93.73 ± 0.60	92.83 ± 0.76	93.26 ± 0.40	0.87 ± 0.00	0.98 ± 0.00	g = 0.0005, c = 1, j = 1
DPC	0	93.79 ± 1.12	95.68 ± 0.78	94.84 ± 0.72	0.90 ± 0.01	0.99 ± 0.01	g = 0.0005, c = 1, j = 2
N5 DPC	−0.1	83.42 ± 1.77	87.73 ± 2.00	85.69 ± 1.10	0.71 ± 0.02	0.93 ± 0.00	g = 1e−05, c = 9, j = 1
C5 DPC	−0.1	90.21 ± 0.91	93.62 ± 0.96	92.00 ± 0.50	0.84 ± 0.01	0.97 ± 0.00	g = 0.0005, c = 1, j = 2
N5C5 DPC	−0.2	93.60 ± 0.72	92.67 ± 1.16	93.11 ± 0.70	0.86 ± 0.01	0.98 ± 0.00	g = 0.0001, c = 1, j = 1
N5 bin	−0.1	86.91 ± 0.82	88.81 ± 1.47	87.91 ± 0.73	0.76 ± 0.01	0.94 ± 0.00	g = 0.5, c = 2, j = 1
C5 bin	−0.2	91.18 ± 0.92	86.61 ± 1.68	88.80 ± 1.14	0.78 ± 0.02	0.96 ± 0.00	g = 0.5, c = 1, j = 2
N5C5 bin	0.2	89.20 ± 1.11	91.14 ± 1.61	90.22 ± 1.05	0.80 ± 0.02	0.96 ± 0.00	g = 0.05, c = 1, j = 4
N10 bin	−0.2	86.39 ± 2.73	89.68 ± 1.79	88.42 ± 1.05	0.76 ± 0.02	0.94 ± 0.01	g = 0.1, c = 2, j = 2
C10 bin	−0.2	79.87 ± 2.30	86.49 ± 2.43	83.96 ± 1.91	0.66 ± 0.03	0.90 ± 0.01	g = 0.05, c = 3, j = 1
N10C10 bin	−0.4	86.89 ± 2.70	91.62 ± 2.92	89.83 ± 1.31	0.79 ± 0.02	0.96 ± 0.00	g = 0.1, c = 1, j = 1
AAC + motif	−0.1	95.51 ± 0.86	95.35 ± 0.85	95.42 ± 0.77	0.91 ± 0.01	0.99 ± 0.00	g = 0.001, c = 6, j = 1
DPC + motif	0	94.15 ± 0.92	96.94 ± 0.49	95.71 ± 0.38	0.91 ± 0.00	0.99 ± 0.00	g = 0.0005, c = 1, j = 2

This table shows average performance (mean ± standard deviation) of models on randomly generated training datasets (bagging)

MCC Matthews correlation coefficient, *AAC* amino acid composition, *DPC* dipeptide composition, *N5* first 5 residues from N terminus, *C5* first 5 residues from C terminus, *N5C5* first 5 residues from N and C terminus respectively, *bin* binary profile, *AAC + motif* amino acid composition with MERCI motif score, *DPC + motif* dipeptide composition with MERCI motif score, *SVM parameters g* gamma parameter of the radial basis function, *c* trade-off between training error and margin, *j* regularization parameter (cost-factor, by which training errors on positive examples outweigh errors on negative examples)

0.80 while the N10C10 bin model performed the best among the 10-residue position models (accuracy 89.83% and MCC 0.79).

Hybrid model

In the previous motif analysis, motifs exclusive to the A-cell epitopes were identified using the MERCI program. We checked whether the motif information added to composition could help improve the performance of prediction. Indeed, the AAC + motif model that combined the information of presence of motifs exclusive to A-cell epitopes with amino acid composition achieved a better performance than AAC model on the training–testing dataset giving an accuracy value of 95.42% and MCC value of 0.91 (Table 1). Yet, the best model among all the feature combinations was that of motif information combined to the DPC (DPC + motif model) that gave an accuracy of 95.71% and the MCC value 0.91.

Performance of the models on the independent datasets

The independent dataset was generated in the same way as training dataset (by sampling) and consisted of ~ 20% of the total dataset resulting in 61 positive sequences and 77 negative sequences. The average performances of the AAC and DPC models on the independent dataset were comparable to those on the training–testing dataset with AAC model giving an accuracy of 93.91% and MCC of 0.88 while DPC model giving an accuracy of 94.64% and MCC of 0.89 (Table 2). The MCC values of the N5C5 AAC and N5C5 DPC models in the independent dataset evaluation (0.88 and 0.88 respectively) were also close to those found in the training–testing dataset evaluation. Similar to the training–testing dataset results, the N5C5 bin and N10C10 bin models performed the best (MCC 0.82 and 0.81 respectively) among the binary models on the independent dataset. The hybrid model (AAC + motif and DPC + motif) gave accuracies 94.35 and 95.00% respectively, while the MCC values found were 0.89 and 0.90 respectively when evaluated on the independent dataset (Table 2). Figure 5 is a plot of the MCC values of various SVM models on the training–testing dataset along with the MCC values of the models obtained on the independent dataset drawn shown as bars. The MCC values on both the datasets are comparable for each of the models developed indicating that the models are not over optimized on the training–testing dataset.

Alternate negative dataset of random peptides

In the absence of the experimentally verified non-epitopes, the human endogenous circulating peptides were considered to be negative sequences. There are two major issues with this dataset; firstly, it is possible that some of endogenous circulating peptides are A-cell epitopes and secondly, the size of negative dataset is

Table 2 The performance of SVM-based models developed using various features; models were evaluated on independent dataset (external cross-validation)

Features	Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	AUROC	Parameters
AAC*	− 0.1	94.10 ± 2.70	93.77 ± 3.28	93.91 ± 2.00	0.88 ± 0.03	0.98 ± 0.00	g = 0.001, c = 3, j = 3
N5 AAC	0	89.75 ± 3.61	90.88 ± 3.67	90.32 ± 2.27	0.81 ± 0.04	0.95 ± 0.01	g = 0.0005, c = 2, j = 1
C5 AAC	0	91.12 ± 3.53	91.64 ± 2.90	91.40 ± 2.13	0.83 ± 0.04	0.97 ± 0.01	g = 0.001, c = 9, j = 1
N5C5 AAC	− 0.2	94.61 ± 3.35	93.59 ± 3.23	94.07 ± 2.16	0.88 ± 0.04	0.98 ± 0.00	g = 0.0005, c = 1, j = 1
DPC	0	93.77 ± 2.76	95.32 ± 1.40	94.64 ± 1.24	0.89 ± 0.02	0.99 ± 0.00	g = 0.0005, c = 1, j = 2
N5 DPC	− 0.1	81.68 ± 4.25	87.36 ± 2.78	84.62 ± 2.65	0.69 ± 0.05	0.93 ± 0.01	g = 1e−05, c = 9, j = 1
C5 DPC	− 0.1	92.31 ± 3.36	94.71 ± 2.45	93.55 ± 1.40	0.87 ± 0.02	0.98 ± 0.01	g = 0.0005, c = 1, j = 2
N5C5 DPC	− 0.2	94.10 ± 3.06	93.75 ± 2.33	93.90 ± 1.49	0.88 ± 0.03	0.98 ± 0.01	g = 0.0001, c = 1, j = 1
N5 bin	− 0.1	88.46 ± 2.90	89.43 ± 3.32	88.98 ± 2.36	0.78 ± 0.04	0.95 ± 0.01	g = 0.5, c = 2, j = 1
C5 bin	− 0.2	93.70 ± 3.03	87.88 ± 4.25	90.63 ± 2.43	0.82 ± 0.04	0.97 ± 0.01	g = 0.5, c = 1, j = 2
N5C5 bin	0.2	90.95 ± 3.18	91.13 ± 3.44	91.03 ± 2.77	0.82 ± 0.05	0.97 ± 0.01	g = 0.05, c = 1, j = 4
N10 bin	− 0.2	89.38 ± 6.68	90.46 ± 4.67	90.01 ± 3.26	0.79 ± 0.06	0.95 ± 0.03	g = 0.1, c = 2, j = 2
C10 bin	− 0.2	85.02 ± 8.02	85.24 ± 5.15	85.19 ± 4.09	0.69 ± 0.09	0.93 ± 0.03	g = 0.05, c = 3, j = 1
N10C10 bin	− 0.4	88.73 ± 5.95	92.33 ± 5.69	91.04 ± 2.52	0.81 ± 0.05	0.97 ± 0.02	g = 0.1, c = 1, j = 1
AAC + motif	− 0.1	93.11 ± 1.86	95.33 ± 3.13	94.35 ± 1.67	0.89 ± 0.03	0.99 ± 0.00	g = 0.001, c = 6, j = 1
DPC + motif	0	93.28 ± 2.38	96.36 ± 1.70	95.00 ± 1.25	0.90 ± 0.02	0.99 ± 0.00	g = 0.0005, c = 1, j = 2

This table shows average performance (mean ± standard deviation) of models on randomly generated independent datasets (bagging)

MCC Matthews correlation coefficient, *AAC* amino acid composition, *DPC* dipeptide composition, *N5* first 5 residues from N terminus, *C5* first 5 residues from C terminus, *N5C5* first 5 residues from N and C terminus respectively, *bin* binary profile, *AAC + motif* amino acid composition with MERCI motif score, *DPC + motif* dipeptide composition with MERCI motif score, *SVM parameters g* gamma parameter of the radial basis function, , *c* trade-off between training error and margin, *j* regularization parameter (cost-factor, by which training errors on positive examples outweigh errors on negative examples)

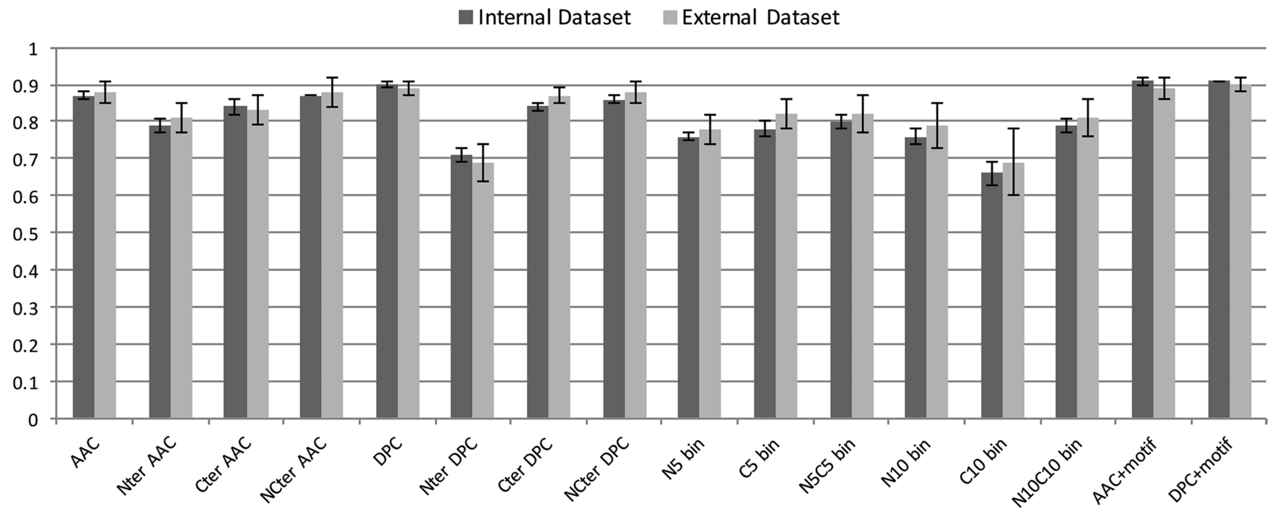

Fig. 5 Comparison of the support vector machine-based prediction models on the training–testing and the independent datasets. The striped bars correspond to the Matthews correlation coefficient (MCC) values obtained for the models on the training–testing dataset and the solid line joins the MCC values of the models on the independent dataset. For each model, the MCC values for the training–testing dataset and the independent dataset are comparable indicating the reliable prediction capabilities of the models

small. In this study, we developed models on alternate dataset. In alternate dataset, non-epitopes were derived randomly from human proteins available in the Swiss-Prot database. Our alternate dataset the non-epitopes (or random peptide) 10 times the number of A-cell epitopes. Bagging procedure of sampling was performed to create ten training and ten independent datasets for internal and external validation respectively. The DPC and DPC + motif models performed better than other models on the main dataset. As shown in Table 3, the

Table 3 The average performance of A-cell epitope prediction models on training and independent dataset

Features	Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	Parameters
Internal validation: performance on training dataset, evaluated using fivefold cross-validation						
DPC	− 0.2	87.49 ± 1.41	98.70 ± 0.16	97.68 ± 0.22	0.86 ± 0.01	g: 0.0005, c: 1, j: 4
DPC + Motif	− 0.2	87.81 ± 1.01	99.30 ± 0.10	98.25 ± 0.17	0.89 ± 0.01	g: 0.0005, c: 1, j: 4
External validation: performance on independent dataset						
DPC	− 0.2	87.54 ± 4.31	98.87 ± 0.28	97.84 ± 0.41	0.87 ± 0.02	g: 0.0005, c: 1, j: 4
DPC + Motif	− 0.2	77.86 ± 5.84	99.28 ± 0.30	97.33 ± 0.58	0.83 ± 0.04	g: 0.0005, c: 1, j: 4

These training and independent datasets were created from alternate datasets using bagging. In alternate dataset, negative or non-epitopes were derived from human proteins. The performance values have been reported as mean ± standard deviation for each model

MCC Matthews correlation coefficient, *DPC* dipeptide composition, *DPC + motif* dipeptide composition with MERCI motif score, *SVM parameters g* gamma parameter of the radial basis function, *c* trade-off between training error and margin, *j* regularization parameter (cost-factor, by which training errors on positive examples outweigh errors on negative examples)

performance of models on alternate dataset is similar to the performance of models on main dataset. Overall, the performance values are comparable indicating that the choice of negative dataset in the absence of the experimentally verified negative sequences or the total number of training examples has a minimal effect on the performance of the prediction models developed in this study.

VaxinPAD: the web interface for prediction of A-cell epitopes

For providing the SVM-based A-cell epitope prediction methods developed in the present study to the scientific community, we designed an in silico platform VaxinPAD (available at http://webs.iiitd.edu.in/raghava/vaxinpad/). The platform has a couple of utilities that may help the user in designing peptide-based adjuvants as well enhance or diminish the immunomodulatory potential of the query sequence.

Designing of vaccine adjuvants

The 'PREDICTION' module of the VaxinPAD platform allows the user to check whether the query sequence would be immunomodulatory on the basis of SVM score. It allows the virtual screening of the A-cell epitopes among a library of input peptide sequences.

Designing analogs of adjuvant peptides

The 'ANALOGS' module of VaxinPAD enables a user to generate all possible single residue position substituents of a query peptide sequence and predict potential immunomodulators among the analogs generated.

Immunomodulatory regions in a protein

The 'PROTEIN ADJUVANTS' module of VaxinPAD does a window search across the length of a query protein sequence to identify immunomodulatory patches, the window size being user-defined. LL37, a well-known immunomodulatory peptide is a 37 amino acid peptide derived from human Cathelicidin. This menu may help the researchers in identifying more such peptides that are immunomodulatory.

Peptide sequences dataset

Finally, VaxinPAD includes a menu 'DATASET' that includes a list of immunomodulatory peptides collected from literature. Among the sequences in the database only the peptides of length 3–30 were used for development of prediction models in the current work.

Discussion

Previously, peptide-based vaccine adjuvants were largely being developed as ligands of innate immunity receptors like TLR-4 and TLR-2 [45] or as self-assembling nanostructure forming entity [46]. Recently, it has been realized that short immunomodulatory peptides can be developed as potential vaccine adjuvants [4]. Cationic host defense peptides were previously known to have antibacterial activity by direct killing of the pathogen [47]. Of late, these peptides have been found to evoke the innate immunity by a variety of mechanisms [6]. A majority of these mechanisms involve pattern recognition receptors (PRRs) playing important roles especially in the antigen presenting cells (APCs) like dendritic cells, macrophages, etc. Since these peptides activate APCs, we call these peptides as 'A-cell epitopes' (antigen presenting cell epitopes). The A-cell epitopes undertaken in the present study were collected from the patent literature that included host defense peptides as immunomodulatory sequences. To the best of the authors' knowledge, the present study is the first attempt to develop an in silico tool for designing innate immunomodulatory peptides as the first step towards engineering novel peptide-based vaccine adjuvants.

An important finding in this study was that the residues preferred in A-cell epitopes include arginine (R). Arginine enrichment of the peptide sequences is an important aspect of increasing the cellular uptake of cell-penetrating peptides (CPPs) [48]. Cathelicidins are recognized as an important class of host defense peptides that includes many arginine-rich peptides [49]. Further, human cathelicidin-derived peptide LL37 that is rich in basic residues arginine and lysine has been reviewed as a promising immunomodulatory peptide with cell penetrating properties [50]. Hence, sequence analysis of the A-cell epitopes may indicate cell-penetrating ability to be an associated property of the A-cell epitopes.

Another aspect of our sequence analysis is the occurrence of *n*-mers found sparsely in the naturally occurring proteins. Patel et al. [51] found that peptide pentapeptides occurring rarely in the universal proteome when introduced into the end of the antigenic sequence enhanced its antigenicity and also suggested that on exogenous addition these rare pentapeptides could act as immunomodulators and thus could be developed as adjuvants. In our analysis too, we found tripeptides, tetrapeptides and pentapeptides occurring rarely in Swiss-Prot proteins to be present more in the A-cell epitopes than non-epitopes. This fits well with the intuition that the immune system is more likely to react to rarely encountered sequence motifs than frequent ones.

On evaluating the performance of SVM models based on composition, the dipeptide composition showed no improvement over amino acid composition. The binary models also showed a lower performance than composition-based models. On the other hand, addition of the motif information increased the performance of both the amino acid composition model as well as dipeptide composition to achieve the maximum accuracies of ~ 96%. In our study, we implemented our best models- the dipeptide composition-based and that based on the hybrid of dipeptide composition and motifs. These models were better than other models possibly owing to the fact that the dipeptide composition provides more information in comparison to simple composition. Dipeptide composition provides the information about the amino acid fraction as well as their local order [34]. It is well known fact that there are certain patterns/motifs present in the proteins/peptides which are responsible for its biological activity [52–55]. In past also, many methods have been developed which have shown that adding the motif information increases the accuracy of model [56, 57]. In our analysis, we have identified certain motifs which are exclusively present in A-cell epitopes. We also observed that adding the information of these motifs with dipeptide composition improved the accuracy of the model. Therefore, we believe that this model will help in classifying the A-cell epitope from non-A cell epitope more accurately in comparison to other models.

We also checked whether the random distribution of the main dataset into the dataset used for training the models (internal dataset) and that evaluating them independently (external dataset) renders a bias in the performance of prediction models. Further, the choice of sequences assumed to be the non-epitopes could also affect the performance of the prediction models. A low number of positive sequences in the total dataset could be a third source of influence on the robustness of prediction models. Using various methods, we observed that all of these three factors have a negligible effect on the performance of the best performing SVM-based prediction models for A-cell epitopes.

The peptides designed using the tools developed in the present study might act by various mechanisms and receptors for activating the innate immunity owing to the fact that the training dataset of the prediction models contains peptides acting by diverse cell signaling routes. Hence, the in silico tool presented here could help an investigator to begin with a choice of peptides that may be the starting molecules for the development of vaccine adjuvants. However, there are certain limitations associated with this model, for example, the method does not consider modifications (e.g., post-translational modifications) and other topological aspects during model development. Secondly, whether the predicted peptides actually prove to be useful as adjuvants, would have to be tested experimentally. Another limiting aspect of the present study is the exclusion of very long immunomodulatory sequences and lastly, the size of the dataset used in the study. The model can be further optimized by incorporating more peptide features such as physico–chemical properties, modifications, etc. Also, with larger datasets and receptor-specific ligands made available in future, studies subsequent to the present investigation might help design peptides eliciting a specific desired innate immune response leading to adjuvancy. Nonetheless, VaxinPAD developed in this study for predicting the immunomodulatory peptides sets a stage for the advancement of rational peptide-based vaccine adjuvant designing.

In silico methods for predicting and identifying DNA and RNA-based immunomodulatory molecules have already been developed [13, 14]. The current study is the first attempt to develop models for predicting immunomodulatory peptides for the development of vaccine adjuvants. Though other biomolecules like lipopolysaccharides and glycosaminoglycan's also cause activation of innate immunity by binding to the PRRs, the literature currently does not hold a sufficient number of molecules for developing prediction models in these categories. Future studies might focus on the development of in silico tools for predicting such immunomodulatory biomolecules for obtaining new vaccine adjuvants. In addition to this, peptides with non-natural chemical modifications might offer better adjuvants too. Correspondingly, computational tools for prediction of modified peptides might also become an area of development.

Conclusion

Host Defense Peptides have been realized as promising immunomodulators likely to become potential vaccine adjuvants [47]. With immunomodulatory properties, novel peptides predicted to be A-cell epitopes using the models developed here are also likely to have potential to provide host protection against pathogens. Many host defense peptides (HDPs) with known immunomodulatory effects are already in clinical trials [47]. Despite the associated toxicity of the A-cell epitopes due to their pleiotropic effects on the immune system, rational design of innate defense regulators (IDRs) that are synthetic analogs of HDPs is in pressing demand for having immunopotentiators with reduced toxicity and increased specificity of immune responses. We have developed SVM-based models for prediction of A-cell epitopes that could be used to formulate vaccine adjuvants. These models have been implemented in the form of webserver VaxinPAD available at http://webs.iiitd.edu.in/raghava/vaxinpad/ and http://crdd.osdd.net/raghava/vaxinpad/ freely to the scientific community.

Additional file

Additional file 1: Table S1. Dataset (A-cell epitopes and non-epitope) used for training, testing and validation; including source of information. **Table S2.** The percent amino acid composition and difference in composition for A-cell epitopes, non-epitopes and human proteins including Students t-test with p-value and adjusted p-value. **Table S3.** The percent dipeptide composition and difference in composition in A-cell epitopes, non-epitopes and human proteins. **Table S4.** The percent tripeptide compositions of A-cell epitopes, non-epitopes and human proteins. Rows are sorted in decreasing order of values in the 5th column. **Table S5.** Motifs exclusively found in A-cell epitopes and their frequency of occurrence, sorted in the decreasing order of counts. **Table S6.** Motifs exclusively found in non-epitopes and their frequency of occurrence, sorted in decreasing order of counts. **Table S7.** Performance of various classifiers on the training–testing dataset and the independent dataset. **Table S8.** The performance of SVM-based A-cell epitope prediction models developed using various features; models were evaluating on training and independent datasets using fivefold cross-validation.

Abbreviations

APC: antigen presenting cell; PRR: pattern recognition receptor; HDP: host defense peptide; IDR: innate defense regulator; TLR: Toll-like receptor; CAMP: cathelicidin antimicrobial peptide; MHC: major histocompatibility complex; SVM: support vector machine; MERCI: Motif—EmeRging and with Classes—Identification; TSL: two-sample logo; CPP: cell penetrating peptides.

Authors' contributions

PA and GN collected the data. PA, KC and GN organized the data. KC, GN and PA performed the experiments. GN and KC developed the web interface. GN, KC, PA and GPSR analyzed the data. GN and GPSR prepared the manuscript. GPSR conceived the idea and coordinated the project. All authors read and approved the final manuscript.

Author details

[1] Bioinformatics Centre, Institute of Microbial Technology, Chandigarh 160036, India. [2] Centre for Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India.

Acknowledgements

Authors are thankful to the projects, Open Source Drug Discovery (OSDD) and GENESIS BSC0121. GN and KC are thankful to CSIR and PA is thankful to the Department of Science and Technology (DST), Government of India for providing fellowships.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets used for this study are provided at http://webs.iiitd.edu.in/raghava/vaxinpad/sequences.php and http://crdd.osdd.net/raghava/vaxinpad/sequences.php as well as Additional file 1: Table S1.

Consent for publication

Not required.

Ethics approval and consent to participate

Not required in this study.

Funding

This work was supported by the Council of Scientific and Industrial Research (CSIR), India and the Department of Biotechnology (DBT), Government of India.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 15 January 2018 Accepted: 23 June 2018
Published online: 03 July 2018

References

1. Clem AS. Fundamentals of vaccine immunology. J Glob Infect Dis. 2011;3:73–8. http://www.jgid.org/text.asp?2011/3/1/73/77299. Accessed 16 May 2018.
2. Coffman RL, Sher A, Seder RA. Vaccine adjuvants: putting innate immunity to work. Immunity. 2010;33:492–503.
3. Azmi F, Ahmad Fuaad AAH, Skwarczynski M, Toth I. Recent progress in adjuvant discovery for peptide-based subunit vaccines. Hum Vaccines Immunother. 2014;10:778–96.
4. Dong JC, Kobinger GP. Hypothesis driven development of new adjuvants: short peptides as immunomodulators. Hum Vaccines Immunother. 2013;9:808–11. https://doi.org/10.4161/hv.22972.
5. Zasloff M. Antimicrobial peptides of multicellular organisms. Nature. 2002;415:389–95. https://doi.org/10.1038/415389a.
6. Mansour SC, Pena OM, Hancock REW. Host defense peptides: front-line immunomodulators. Trends Immunol. 2014;35:443–50.
7. Afacan NJ, Yeung ATY, Pena OM, Hancock REW. Therapeutic potential of host defense peptides in antibiotic-resistant infections. Curr Pharm Des. 2012;18:807–19.
8. Litvinov S. New adjuvants for accelerated and enhanced antibody response. Nat Methods. 2009. http://www.nature.com/articles/nmeth.f.279. Accessed 16 May 2018.
9. Torres-Juarez F, Cardenas-Vargas A, Montoya-Rosales A, González-Curiel I, Garcia-Hernandez MH, Enciso-Moreno JA, et al. LL-37 immunomodulatory activity during *Mycobacterium tuberculosis* infection in macrophages. Infect Immun. 2015;83:4495–503. https://doi.org/10.1128/IAI.00936-15.

10. Vandamme D, Landuyt B, Luyten W, Schoofs L. A comprehensive summary of LL-37, the factotum human cathelicidin peptide. Cell Immunol. 2012;280:22–35.
11. Pütsep K, Carlsson G, Boman HG, Andersson M. Deficiency of antibacterial peptides in patients with morbus Kostmann: an observation study. Lancet. 2002;360:1144–9.
12. Dhanda SK, Usmani SS, Agrawal P, Nagpal G, Gautam A, Raghava GPS. Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. Brief Bioinform. 2017;18:467–78. https://doi.org/10.1093/bib/bbw025.
13. Nagpal G, Gupta S, Chaudhary K, Dhanda SK, Prakash S, Raghava GPS. Vaccine DA: Prediction, design and genome-wide screening of oligodeoxynucleotide-based vaccine adjuvants. Sci Rep. 2015;5:12478. http://www.nature.com/articles/srep12478. Accessee 16 May 2018.
14. Chaudhary K, Nagpal G, Dhanda SK, Raghava GPS. Prediction of immunomodulatory potential of an RNA sequence for designing non-toxic siRNAs and RNA-based vaccine adjuvants. Sci Rep. 2016;6:20678. http://www.nature.com/articles/srep20678. Accessed 16 May 2018.
15. Wee LJK, Simarmata D, Kam Y-W, Ng LFP, Tong JC. SVM-based prediction of linear B-cell epitopes using Bayes feature extraction. BMC Genomics. 2010;11(Suppl 4):S21. https://doi.org/10.1186/1471-2164-11-S4-S21.
16. Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. BMC Bioinform. 2010;11:381. https://doi.org/10.1186/1471-2105-11-381.
17. Rubinstein ND, Mayrose I, Pupko T. A machine-learning approach for predicting B-cell epitopes. Mol Immunol. 2009;46:840–7.
18. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 2006;15:2558–67. https://doi.org/10.1110/ps.062405906.
19. Gupta S, Ansari HR, Gautam A, Raghava GPS, Open Source Drug Discovery Consortium GP. Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. Biol Direct. 2013;8:27. https://doi.org/10.1186/1745-6150-8-27.
20. Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS ONE. 2013;8:e62216. https://doi.org/10.1371/journal.pone.0062216.
21. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 2006;65:40–8. https://doi.org/10.1002/prot.21078.
22. Assis LM, Sousa JR, Pinto NFS, Silva AA, Vaz AFM, Andrade PP, et al. B-cell epitopes of antigenic proteins in *Leishmania infantum*: an in silico analysis. Parasite Immunol. 2014;36:313–23. https://doi.org/10.1111/pim.12111 .
23. Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res. 2010;6:6. http://www.immunome-research.com/content/6/1/6. Accessed 16 May 2018.
24. Sette A. The immune epitope database and analysis resource: from vision to blueprint. Genome Inform. 2004;15:299.
25. Zhang H, Wang P, Papangelopoulos N, Xu Y, Sette A, Bourne PE, et al. Limitations of Ab initio predictions of peptide binding to MHC class II molecules. PLoS ONE. 2010;5:e9272. https://doi.org/10.1371/journal.pone.0009272.
26. Reche PA, Glutting J-P, Zhang H, Reinherz EL. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics. 2004;56:405–19. https://doi.org/10.1007/s00251-004-0709-7.
27. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009;10:296. http://www.biomedcentral.com/1471-2105/10/296. Accessed 16 May 2018.
28. Bhasin M, Raghava GPS. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine. 2004;22:3195–204.
29. Dhanda SK, Vir P, Raghava GPS. Designing of interferon-gamma inducing MHC class-II binders. Biol Direct. 2013;8:30. https://doi.org/10.1186/1745-6150-8-30.
30. Lata S, Bhasin M, Raghava GPS. Application of machine learning techniques in predicting MHC binders. Methods Mol Biol. 2007;409:201–15. https://doi.org/10.1007/978-1-60327-118-9_14.
31. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC class-I binding sites. Bioinformatics. 2003;19:1009–14.
32. Hu L, Boos K-S, Ye M, Zou H. Analysis of the endogenous human serum peptides by on-line extraction with restricted-access material and HPLC–MS/MS identification. Talanta. 2014;127:191–5.
33. Tucholska M, Florentinus A, Williams D, Marshall JG. The endogenous peptides of normal human serum extracted from the acetonitrile-insoluble precipitate using modified aqueous buffer with analysis by LC–ESI-Paul ion trap and Qq-TOF. J Proteomics. 2010;73:1254–69.
34. Agrawal P, Bhalla S, Chaudhary K, Kumar R, Sharma M, Raghava GPS. In silico approach for prediction of antifungal peptides. Front Microbiol. 2018;9:323. https://doi.org/10.3389/fmicb.2018.00323/full.
35. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics. 2011;27:1231–8. https://doi.org/10.1093/bioinformatics/btr110.
36. Dhanda SK, Gupta S, Vir P, Raghava GPS. Prediction of IL4 inducing peptides. Clin Dev Immunol. 2013;2013:263952. http://www.hindawi.com/journals/jir/2013/263952/. Accessed 16 May 2018.
37. Tran E, Turcotte S, Gros A, Robbins PF, Lu Y-C, Dudley ME, et al. Cancer immunotherapy based on mutation-specific CD4$^+$ T cells in a patient with epithelial cancer. Science. 2014;344:641–5. https://doi.org/10.1126/science.1251102.
38. Schirmbeck R, Böhm W, Fissolo N, Melber K, Reimann J. Different immunogenicity of H-2 Kb-restricted epitopes in natural variants of the hepatitis B surface antigen. Eur J Immunol. 2003;33:2429–38. https://doi.org/10.1002/eji.200324125.
39. Witten IH, Ian H, Frank E, Hall MA, Mark A, Pal CJ. Data mining: practical machine learning tools and techniques. Burlington: Morgan Kaufmann; 2016.
40. Zhao Y, Pinilla C, Valmori D, Martin R, Simon R. Application of support vector machines for T-cell epitopes prediction. Bioinformatics. 2003;19:1978–84.
41. Bhasin M, Raghava GPS. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics. 2004;20:421–3. https://doi.org/10.1093/bioinformatics/btg424.
42. Schölkopf B, Burges CJC, Smola AJ. Advances in kernel methods: support vector learning. MIT Press; 1999. https://dl.acm.org/citation.cfm?id=299094. Accessed 16 May 2018.
43. Kumar V, Agrawal P, Kumar R, Bhalla S, Usmani SS, Varshney GC, et al. Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues. Front Microbiol. 2018;9:725. https://doi.org/10.3389/fmicb.2018.00725/full.
44. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22:1536–7. https://doi.org/10.1093/bioinformatics/btl151.
45. Shanmugam A, Rajoria S, George AL, Mittelman A, Suriano R, Tiwari RK. Synthetic Toll like receptor-4 (TLR-4) agonist peptides as a novel class of adjuvants. PLoS ONE. 2012;7:e30839. https://doi.org/10.1371/journal.pone.0030839.
46. Rudra JS, Tian YF, Jung JP, Collier JH. A self-assembling peptide acting as an immune adjuvant. Proc Natl Acad Sci. 2010;107:622–7.
47. Hilchie AL, Wuerth K, Hancock REW. Immune modulation by multifaceted cationic host defense (antimicrobial) peptides. Nat Chem Biol. 2013;9:761–8. http://www.nature.com/articles/nchembio.1393. Accessed 16 May 2018.
48. Bechara C, Sagan S. Cell-penetrating peptides: 20 years later, where do we stand? FEBS Lett. 2013;587:1693–702. https://doi.org/10.1016/j.febslet.2013.04.031.
49. Kościuczuk EM, Lisowski P, Jarczak J, Strzałkowska N, Jóźwik A, Horbańczuk J, et al. Cathelicidins: family of antimicrobial peptides: a review. Mol Biol Rep. 2012;39:10957–70.
50. Shakya AK, Kumar A, Holmdahl R, Nandakumar KS. Macrophage-derived reactive oxygen species protects against autoimmune priming with a defined polymeric adjuvant. Immunology. 2016;147:125–32. https://doi.org/10.1111/imm.12546.
51. Patel A, Dong JC, Trost B, Richardson JS, Tohme S, Babiuk S, et al. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. PLoS ONE. 2012;7:e43802. https://doi.org/10.1371/journal.pone.0043802.
52. Fink JS, Verhave M, Kasper S, Tsukada T, Mandel G, Goodman RH. The CGTCA sequence motif is essential for biological activity of the vasoactive

intestinal peptide gene cAMP-regulated enhancer. Proc Natl Acad Sci USA. 1988;85:6662–6.

53. McGowan S, O'Connor JR, Cheung JK, Rood JI. The SKHR motif is required for biological function of the VirR response regulator from *Clostridium perfringens*. J Bacteriol. 2003;185:6205–8.
54. Ono H, Kozmik Z, Yu J-K, Wada H. A novel N-terminal motif is responsible for the evolution of neural crest-specific gene-regulatory activity in vertebrate FoxD3. Dev Biol. 2014;385:396–404.
55. Wakasugi K, Schimmel P. Highly differentiated motifs responsible for two cytokine activities of a split human tRNA synthetase. J Biol Chem. 1999;274:23155–9.
56. Chaudhary K, Kumar R, Singh S, Tuknait A, Gautam A, Mathur D, et al. A web server and mobile app for computing hemolytic potency of peptides. Sci Rep. 2016;6:22843. http://www.nature.com/articles/srep22843. Accessed 16 May 2018.
57. Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, Delattre O, et al. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic Acids Res. 2010;38:e126. https://doi.org/10.1093/nar/gkq217.